Reasoning with data models and modeling in the big data era Editors Colophon

Book · February 2024

DOI: 10.17619/UNIPB/1-1815

4 authors:

Susanne Podworny
Paderborn University
70 PUBLICATIONS **314** CITATIONS
SEE PROFILE

Daniel Frischemeier
University of Münster
119 PUBLICATIONS **556** CITATIONS
SEE PROFILE

Michal Dvir
University of Haifa
14 PUBLICATIONS **89** CITATIONS
SEE PROFILE

Dani Ben-Zvi
University of Haifa
180 PUBLICATIONS **6,233** CITATIONS
SEE PROFILE

Minerva School 2022

Reasoning

with data models and modeling in the big data era

Editors

Susanne Podworny
Daniel Frischemeier
Michal Dvir
Dani Ben-Zvi

UNIVERSITÄT PADERBORN

Podworny, S., Frischemeier, D., Dvir, M. & Ben-Zvi, D., Eds. 2024.
Minerva School 2022:
Reasoning with data models and modeling in the big data era.
https://doi.org/10.17619/UNIPB/1-1815

This project was funded through a generous grant from the Minerva Stiftung.

Design and layout:
Tim Erickson, eeps media,
Oakland, California USA

Colophon

This book was formatted in the Apple ecosystem, using Adobe InDesign for typesetting and layout. The body text is Arno Pro, and heads are in Avenir Next Condensed. We used CODAP, R, Illustrator, and Keynote for graphics and EquationMaker and LaTeX for equations.

Contents

Preface

In an era increasingly driven by the transformative power of data, the significance of data and data modeling cannot be overstated. This book explores these pivotal topics in the context of data science and education, highlighting their critical role in advancing our understanding of complex phenomena to meet diverse educational needs in the big data era.

It is with immense pleasure that I present this scholarly volume, a collaborative effort that brings together young and experienced scientists from different disciplines to facilitate an interdisciplinary discourse. The contributors to this compilation actively participated in our 2022 Minerva School "An Interdisciplinary Exploration of Future Data Pedagogies and Digital Tools to Nurture Citizens' Reasoning with Models and Modeling in the Big Data Era" held in Paderborn, Germany. This Minerva School brought together a range of expertise from different subject domains, theoretical perspectives, and methodological approaches.

The chapters in this volume emanate from the presentations and enriching discussions that characterized the Minerva School. This ambitious undertaking wouldn't have been possible without the support of the Minerva Foundation, whose financial backing not only materialized this publication but also facilitated the seamless execution of the conference.

My heartfelt gratitude extends to all of the attendees of the conference, who generously shared their research and actively participated in the interdisciplinary dialogues. Your invaluable contributions have enriched the Minerva School and reflect the collaborative spirit that defines our scientific community.

To the authors who have generously contributed to this volume, my profound thanks. Your commitment to expanding the boundaries of knowledge and sharing your expertise has undeniably shaped the essence of this book, illuminating new trajectories for the future. The variety of perspectives and the depth of insight provided by each contributor have made this collection a valuable and informative resource.

I extend my deepest appreciation to my co-editors and co-organizers Michal Dvir, Daniel Frischemeier and Dani Ben-Zvi. Your steadfast commitment and collaborative efforts have been instrumental in bringing this project to fruition, and for that, I am profoundly grateful. You are the very best!

Special commendation is reserved for Tim Erickson for his outstanding contributions to the editing, formatting, layout, and presentation of this book. His meticulous attention to detail has significantly enhanced its readability and overall quality. It has been both a pleasure and a privilege working alongside you. (And I'm a little proud that you learned a new word from me).

To all those involved in realizing this endeavor, I extend my sincere appreciation, and I earnestly hope that readers discover in this collection a valuable resource for exploration and discovery.

Susanne Podworny
February 2024, Paderborn University

The multidimensional pedagogical potential of data modeling

Michal Dvir, Susanne Podworny, Dani Ben-Zvi, and Daniel Frischemeier

Data have permeated our everyday lives and has become an indispensable commodity. New forms of data, data visualization, and human interaction with data are changing radically and rapidly. As a result, what it means to be data literate is also changing. Industrial processes, marketing processes, economic processes, and monitoring processes in politics are all based on data and statistics (Gould, 2017; Engel, 2017). The internet as well as sensors and apps have made large and messy datasets available to the public, allowing easy access to meaningful explorations in these domains. However, unscrupulous data manipulation and biased reports have also become a staple of today's reality, empowering a self-selected few to dangerously influence public opinion and policy. To counter these dangers in the big data era, people should not simply be passive recipients of data-based reports, but rather become active data explorers who can plan for, acquire, manage, analyze, model, and infer from data to make informed data-based decisions and judgments. The latter necessitate developing data related skills such as understanding how data can be used to describe and model the world, and critically evaluate the use of data analysis, modeling, and visualization.

This current data-era reality has encouraged educators across various fields to address these new demands. A shared trend has been to move away from the traditional focus on procedures relevant to a discipline (be it mathematics, science, computer science, etc.) and focus more on engaging learners in practices that are authentic to that discipline, to facilitate learning through experience (Garfield & Ben-Zvi, 2008). Even though disciplines differ in their practices, one key practice—*modeling*—appears to be a shared focus (Pfannkuch, Ben-Zvi, & Budgett, 2018). Examining its pedagogical potential is becoming central across various disciplines, resulting in innovative, albeit local, insights. In light of this, the goal of the German-Israeli Minerva School held in August 2022 at Paderborn, Germany, was to facilitate a cross-disciplinary discussion to coalesce these local understandings into one encompassing framework that is deeper and more comprehensive than its individual current strands.

This collection of contributions introduces various perspectives of the Minerva School 2022 participants on the theme of the School: "An interdisciplinary exploration of future data pedagogies and digital tools to nurture citizens' reasoning with models and modeling in the big data era."

This preliminary chapter will introduce some of the perspectives on data models and modeling that arose in this interdisciplinary discussion as represented by the subsequent chapters. We then offer an initial first-pass multidimensional framework derived from the discussion, describing the pedagogical potential of data modeling. Finally, we will use that framework to introduce the chapters themselves.

Data models and modeling: objects, practices, and pedagogies

In this section we provide background for the key terms this book focuses on: *data models*, the practice or process of *data modeling*, and data modeling pedagogies. We start by explaining the definitions of data models and modeling that were adopted in this book. These definitions bridge different disciplinary discourses and showcase the multifaceted or multidimensional nature of data modeling. We then provide some background about data modeling pedagogies, adding an additional dimension to account for the pedagogical potential of data modeling.

Data models and data modeling

Modeling is an inseparable aspect of data handling, understanding, and skills. The term "model" fundamentally means a representation, an analogy, with a descriptive, explanatory, or predictive purpose (Hesse, 1962). The model is typically a simplification of a more complex phenomenon, consisting of a representation of specific elements and possible relations between them (Lesh & Doerr, 2003). As a representation may be found to be

ill-suited for its intended purpose, the model should be constantly evaluated and refined (Hesse, 1962). Modeling refers to the process of model creation, evaluation, and refinement.

Using this broad view, we define a "data model" as a purposeful representation using data. According to this view, because data are a simplified representation of a real-world phenomenon, the data values themselves can serve as models (e.g., Podworny & Frischemeier, **p 15**). Building on that foundation, we can think of "public-friendly" visual or verbal data representations as data models (e.g., Binder, **p 23**; Bielik, **p 33**; Büscher, **p 49**; Engel, **p 101**; Gafny & Ben-Zvi, **p 69**; Gal, **p 91**); as well as more formal models typically represented mathematically (e.g., Gould 2024 **p 81**; Hagenkötter et al., **p 41**).

If data are models, the process of collecting data and analyzing them is a modeling process. This is what many have considered to be "data modeling" (e.g., Hancock, Kaput, & Goldsmith, 1992; Lehrer & Romberg, 1996; Lehrer & Schauble, 2000). Following our definition of data *model*, the data *modeling* process should include the original choices made (such as what attributes should be collected); subsequent refinements such as changing or adding new attributes (Manor & Ben-Zvi, 2015); cleaning the data or tinkering with the data structure (Erickson et al., 2019); constructing data representations; and producing summaries of patterns and variation observed in these representations (Dvir & Ben-Zvi, 2023; Binder 2024, **p 23**). This broad definition includes more obvious and restricted examples of data modeling such as fitting a least-squares line to data or asserting that two variables are independent.

As you will see shortly, the breadth of this definition helps us develop our framework within a deeper and more comprehensive interdisciplinary discussion. We will create a set of what we call *dimensions* in our framework, offering a structure and a common ground for talking about reasoning with data models and modeling. For example, putting data at the center of both the object (the model) and the practice (modeling) highlights that reasoning with data and *data knowledge* (e.g., creating useful data representations) plays an important role—and can prove to be a significant hindrance—when engaging in data modeling (Konold & Higgins, 2003). The emphasis on representations highlights an additional dimension of reasoning with data models and modeling relating to the subject that is being represented, the phenomenon, or the *context*, that is key in evaluating the usefulness of the data model (Wild & Pfannkuch, 1999) and interpreting its meaning. The latter can be particularly challenging when the modeler or data analyst has limited disciplinary (e.g., scientific) knowledge about the modeled phenomenon (Finzer, 2013). An additional important dimension of reasoning with data models and modeling relates to more general modeling skills such as fitting a model and evaluating the extent of the explanation, description, or prediction, that a specific data model provides for a given data set. The latter is an important *driver* of the data modeling process, as it determines whether the current model is sufficient (Dvir & Ben-Zvi, 2018), alongside additional general modeling drivers such as the purpose of the model (Hesse, 1962) and an initial conjecture that is often formulated prior to the data modeling process (Budgett & Pfannkuch, 2018).

As with other cross-disciplinary practices (such as forming data-based conclusions), beyond these three dimensions of data-related knowledge, contextual knowledge, and modeling drivers, an additional facet that is intrinsically related to engaging in data modeling is adopting supportive *habits of mind*, e.g., seeking explanation, a critical stance toward data-based claims, flexibility, and creativity (Makar, Bakker, & Ben-Zvi, 2011), as well as adopting modeling related *norms*, e.g., basing models on data (Dvir & Ben-Zvi, 2018), or enhancing or balancing explanatory and predictive potential (Sainani, 2014). This multifaceted and multidimensional nature of data modeling suggests some of the pedagogical benefits of engaging in data modeling activities, inspired and informed by various modeling-centered pedagogies.

Designing pedagogies centered on data modeling

Modeling-centered pedagogies are a shared interest across various data-related fields. In mathematics education, for example, modeling has been prominent in the last decades (e.g., Stillman et al., 2013), and has been at the forefront of pedagogy improvement and implementation discussions. These include examining young modelers' experiences and mathematical insights (e.g., Gravemeijer, 1999), and aspects of modeling-centered activity design (e.g., model eliciting activities: Lesh, Hoover, & Kelly, 1992). In science education, modeling is considered both a key scientific and engineering practice as well as a crosscutting concept that should be integrated into standards of curriculum, instruction, and assessment, to support students' meaningful learning (National Research Council, 2012). In particular, "data modeling" has been described in a manner that is closely connected to the process of scientific inquiry, involving iterative cycles of posing questions, generating and selecting attributes that can be measured, constructing measures and data representations, and making inferences (Lehrer & Romberg, 1996).

Wild and Pfannkuch (1999) describe a similar investigative cycle as one of the main dimensions of expert statistical reasoning, and they describe modeling as one of the general thinking types expert statisticians employ. The pedagogical potential of modeling has gradually

become the focus of the statistics education community as a means to enculturate learners into statistical practice (Pfannkuch et al., 2018). Engaging in modeling-based activities can support developing learners' statistical reasoning with informal statistical inference, uncertainty, context, data and distribution, and variability (Garfield & Ben-Zvi, 2008). In many cases the modeling-based task centers on representing data (Pfannkuch et al., 2018), thus can be considered as a *data modeling* activity. This can happen in a number of ways. For example, Ben-Zvi, Gravemeijer, and Ainley (2018) advocate for the use of real or realistic data as part of the task design and suggest that we design the task to center on big ideas, incorporating assessment aligned with this focus and a supportive, often digital tool. The growing availability of authentic, large datasets and innovative digital tools can further facilitate students' engagement in modeling tasks and provide educators and activity designers with a wider array of modeling-centered pedagogical opportunities.

Advancements and higher accessibility of new digital tools support innovative data modeling task designs (Biehler et al., 2013), helping learners and designers place more emphasis on key concepts and deep understanding of data-related notions, rather than on procedures and technical calculations (Cobb, 2007). Dynamic visualizations further support reasoning and modeling as they help students examine the underlying mechanisms of phenomena and the models representing them (Rubin & Hammerman, 2006). Furthermore, current digital educational tools (e.g., CODAP, iNZight, TinkerPlots) can be specifically tailored to meet young learners' needs and support their gradual construction of data and modeling concepts previously considered too complicated or out of their reach. Applets and web applications can also facilitate data explorations, providing easy access for interested citizens and instructional designers. More professional tools like R or programming environments like Jupyter notebooks offer more features to explore big data, but may require more support to engage users. The abundance and diversity of digital tools therefore requires an understanding of each tool's unique affordances and limitations, to fruitfully engage learners in activities that will foster their reasoning with data modeling.

While the task itself and the choice of technological tools are important elements of the activity design, additional design aspects also warrant consideration to nurture a productive classroom culture (Ben-Zvi et al., 2018). These include fostering collaboration or discussion norms as well as considering the role of the teacher and the scaffolding they provide, e.g., prompts and questions (Makar et al., 2011). Together, all these aspects of design form add to the multidimensional nature of data modeling pedagogies.

The multidimensional nature of data modeling-centered pedagogies

The definition of data model, as well as the characteristics of data modeling, can vary both within a single discipline (e.g., Schulte, **p 61**) and across disciplines (e.g., Hagenkötter et al., **p 41**, and Gafny and Ben-Zvi, **p 69**). Put together, these different views highlight the multidimensional nature of the data modeling practice, and even more so, of the pedagogies it can inspire. Based on and inspired by Makar et al.'s (2011) framework describing the different elements that support reasoning with informal inference, we offer an initial framework (Table 1) that describes dimensions that characterize, mediate, and foster learners' reasoning with data models and modeling. We hope it can be used to classify various data-modeling-centered pedagogies and related research.

Makar et al. (2011) describe five categories in their framework: *Statistical knowledge, Contextual knowledge, Norms and habits, Inquiry drivers, and Design elements*. We adapted these to the context of data modeling; our framework describes five different dimensions of data modeling pedagogies. Specifically, the *Statistical knowledge category* became the *Data knowledge dimension* and the *Inquiry drivers category* became the *Modeling drivers dimension*. We use the term *dimension* rather than *category* to advocate for their concurrent consideration, despite the distinctions between them. We also adapted some of the elements detailed within each category, e.g., substituting the *Belief* element of the category *Inquiry drivers*, with *Purpose* as an element of the *Modeling drivers* dimension. Two considerations guided this adaptation, relating to (1) prior literature on data models, data modeling, and data modeling-centered pedagogies; and (2) the key characteristics of the new contributions introduced in each chapter of this book. Thus, the resulting framework (Table 1) is not necessarily an exhaustive one, but it lets us highlight the multidimensional nature of the pedagogical potential of data modeling, as well as introduce and classify the work presented in each of the book's chapters. The result of the classification is also noted in Table 1, and will be elaborated in the next subsection.

Dimension	Elements included in the dimension	Examples	Chapters that relate to the element (**explicit**/**implicit**)
Data Knowledge	Data related concepts	Data, variation, distribution, signal and noise, models	**Podworny & Frischemeier p 7** **Binder p 23, Gould p 81,** **Gal p 91**
	Data practices	Data collection, data cleaning, data moves (Erickson et al., 2019), data visualizations & representations (e.g., plots, graphs, tables, trees), model fitting, testing, and training	**Podworny & Frischemeier p 7** **Binder p 23, Bielik p 33** **Hagenkötter et al. p 41** **Gafny & Ben-Zvi p 69** **Engel p 101**
Contextual (disciplinary) knowledge	Knowledge about the problem context	Familiarity with aspects of the investigated phenomenon, awareness of possible relationships between elements of the phenomenon	**Bielik p 33** **Hagenkötter et al. p 41** **Büscher p 49, Schulte p 61**
	Disciplinary practices	Disciplinary-specific practices and the values, purpose and endorsed narratives that guide their implementation	**Bielik p 33** **Hagenkötter et al. p 41** **Schulte p 61** **Gal p 91**
Norms and habits	Modeling norms	Basing models on data, balancing explanatory and predictive potential	**Hagenkötter et al. p 41** **Büscher p 49**
	Habits of mind	Seeking explanations, critical stance toward data-based claims, flexibility and creativity	**Podworny & Frischemeier p 7** **Büscher p 49** **Gal p 91** **Engel p 101**
Modeling drivers	Purpose	Explanatory, descriptive, or predictive	**Podworny & Frischemeier p 7** **Büscher p 49, Schulte p 61** **Gafny & Ben-Zvi p 69** **Gould p 81, Gal p 91** **Engel p 101**
	Conjectures	Assumptions based on disciplinary knowledge and prior data	**Gal p 91** **Engel p 101**
	Model fit	Assessing the explanatory potential of the model	**Büscher p 49, Schulte p 61** **Engel p 101**
Design elements	Task	Real or realistic data, focus on central ideas, assessment to monitor and evaluate	**Podworny & Frischemeier p 7,** **Bielik p 33, Büscher p 49** **Gafny & Ben-Zvi p 69** **Gould p 81**
	Tool	CODAP, TinkerPlots, SageModeler purpose-built: cli.math	**Bielik p 33** **Büscher p 49**
	Classroom culture	Role of the teacher, collaboration and discussion norms	**Bielik p 33**

Table 1: The multidimensional framework of data modeling

Using the multidimensional framework to introduce the book's chapters

Before introducing each of the book's chapters, we first note that the work presented in many of the chapters relates to multiple dimensions of this framework, either explicitly or more implicitly. In this introduction, we will limit our descriptions to the elements that each chapter particularly highlights, so that the depiction is concise and the relations and distinctions between each chapter are clearly noticeable. Table 1 provides a summary of the elements identified to be reflected in each of the chapters, along with a distinction between the elements that were more explicit (**in blue**) and those that were implied (**in green**).

The first chapter, by Susanne Podworny and Daniel Frischemeier (**p 15**), describes a study of young learners' perspectives on the concept of data as a model, specifically asking sixth grade students the questions: what are data? And what are data for? In terms of the framework, the chapter particularly relates to *Data related concepts* (the concept of data and its meaning), an element of the *Data knowledge* dimension; and the *Purpose* of models (what are data, as a model, used for) an element of the *Modeling drivers* dimension. This chapter also provides a detailed account of the *Task* (an element of the *Design* dimension) that the students engaged with prior to being asked these two questions, and illustrates that the meaning of even the most basic data related concept, data, might need to be intentionally nurtured, particularly as it can deeply influence learners *Habits of mind* (an element of the *Norms and habits* dimension) when engaging with data and data models

The following chapter, by Karin Binder (**p 23**), extends the discussion on the need to nurture novices' understanding of *Data related concepts* (the *Data knowledge* dimension) by considering more complex concepts such as conditional probabilities. Binder also refers to *Data related practices* as she explores the pedagogical affordances of different visualizations to support novices' (tenth grade students and university students) introduction to this concept. Binder describes the types of reasoning each visualization can mediate, as well as the unique challenges associated with each tool.

While Binder discusses a variety of static, pre-made representations, the chapter by Tom Bielik (**p 33**), focuses on a digital tool (SageModeler) that allows learners to construct their own dynamic representations. The chapter discusses how the digital *Tool* can support students' modeling and *Data practices*, and also highlights additional design elements such as the *Task* itself as well as the *Classroom culture* (e.g., students' collaboration). The chapter also highlights the *Contextual (disciplinary) knowledge* dimension. That is, the modeling task relies heavily on (and can nurture) scientific *Knowledge about the problem context*, and the purpose of the intervention is to develop scientific knowledge as well as scientific *Disciplinary practices* (e.g., computational practices and thinking).

While Bielik focuses on a digital tool, the chapter by Ramona Hagenkötter, Valentina Nachtigall, Katrin Rolka and Nikol Rummel (**p 41**) extends the discussion on the pedagogical potential of data modeling to introduce novices to *Disciplinary practices*, using hands-on experimentation. In contrast to other chapters, the focus is not on the pedagogical potential of data modeling in developing data modeling skills, but rather on mathematical modeling with data. As is the case in Bielik, the pedagogy that is introduced (mathematical hands-on experimentation) underscores elements associated with the *Contextual (disciplinary) knowledge* dimension. The basis of the modeling task is mathematical *Knowledge about the problem context* and the main purpose of the intervention is to develop a wider, more mature, view of mathematical *Disciplinary practices*, complemented by some *Data practices*. Moreover, the goal is also to extend tenth grade students' naïve views of mathematics as a schematic-algorithmic application of procedures, and to introduce them to more mature mathematical *Norms* (the *Norms and habits* dimension).

The *Norms and habits* dimension is also highlighted in the chapter by Christian Büscher (**p 49**). As in the chapter by Hagenkötter et al., this chapter introduces and examines a data-modeling centered pedagogy, with the purpose of nurturing additional (not necessarily data) practices (e.g., argumentation). The activity focuses on the consumption (as opposed to construction) of visual and textual data representations. The activities that the students engage in are intended to gradually foster argumentation *norms* as well as data-related *Habits of mind,* e.g., a critical stance toward data-based claims. The *contextual (disciplinary) knowledge* is an additional dimension that the chapter highlights, explicitly discussing the importance of authentic *Problem contexts*. The latter is a key *Design* element as well, inspiring both the *Task* and the choice of supportive *Tools*. The *Modeling Drivers* dimension is also highlighted in the chapter, as the learners are guided by an authentic *Purpose* of constructing a data-supported argument, based on *Fitting a Model* to given data.

The aspect of the user's *purpose* as an element in the *Modeling drivers* dimension is also highlighted in the chapter by Carsten Schulte (**p 61**). The chapter introduces the traditional view of data models in computer science and computer science education, centering on their representational *purpose*. The creation of the model and its implementation through software highlight the need for *contextual (disciplinary) knowledge*,

including both *knowledge about the problem context* as well as familiarity with *disciplinary practices*. The chapter then suggests some limitations associated with this view, as the *Disciplinary practices* of implementing the model through generating and then running the software change the reality that the model had originally represented, hindering the *Model's fit* to the transformed reality. The chapter concludes with a call to extend the classical view of data modeling in computer science education to include an additional transformative *Purpose*.

The fact that models might have different *purposes* has implications explored in the next chapter, by Ronit Gafny and Dani Ben-Zvi (**p 69**). Like the discussion in the previous chapter, this one also distinguishes between classical data modeling purposes and more modern big data or non-traditional data modeling-centered *Purposes.* The authors explain how the different purpose is often accompanied by different *Data practices*. While the chapter highlights the *modeling drivers* and *data knowledge* dimensions, it also closely relates to the *Design* dimension, as the main product that the chapter offers is an innovative pedagogical approach, illustrated by a sequence of *Tasks* bridging classical and non-traditional *Data practices*, making both more accessible to novices.

The pedagogical potential of classical or "traditional" data modeling to introducing novices to new data practices is also related to the following chapter, by Robert Gould (**p 81**). Focusing on the disciplines of statistics and data science as well, the chapter discusses the properties and pedagogical potential of "traditional" models (*Data related concepts*) as opposed to the focus on data practices in Gafny & Ben-Zvi. The chapter provides many examples of formal traditional models and *Tasks* and explains how they can be used (and may be vital) to introduce novices to important aspects at the core of data science.

The chapter by Iddo Gal (**p 91**) also discusses the question of what data models (and other *Data related concepts*) should be taught. Gal also asks an additional question: taught to whom? This chapter examines these questions in relation to everyday data consumers (rather than Gould's focus on novice data analysts). The chapter suggests distinguishing three different types of consumers, and identifies relevant models for each type (*Data related concepts*), in accordance with the different modeling *Purposes* each type of consumer typically has. The chapter also highlights the need to develop awareness of the *Purpose* of the model and the model's creator, as well as to the assumptions made (*Conjectures*) when generating the model, as part of the *modeling drivers*. The chapter also advocates nurturing the public's *habits of mind* (e.g., developing a critical stance towards data and data representations) to develop and support the latter awareness.

While Gal considers the specific *Data related concepts* that different types of data consumers should be empowered with, the last chapter in the book, by Joachim Engel (**p 101**), extends the discussion on the need for all data consumers to develop a critical stance as part of the *Habits of mind*, and critical awareness to the *purposes* of the model creator (e.g., potential biases), their underlying *conjectures*, and the type and nature of the data that is modeled (and the *Fit* between them). This chapter provides a more holistic discussion of this need, reified by the current and changing information ecosystem characterizing modern life. The chapter identifies key questions that data and data models' consumers should be supported to raise, and concludes by suggesting key ways to nurture their critical appreciation of data and models.

The ten chapters of this book vary in many ways. They reflect different views and insights on data-modeling-centered pedagogies from different disciplinary fields; they consider different types of learners—young, more mature, data professionals, and everyday data consumers; some focus on specific pedagogical implementations or educational data tools, others provide more generalizable blueprints for future data modeling pedagogies, and others provide more holistic perspectives and suggestions; some highlight the need to develop data knowledge and its pedagogical role, others highlight contextual (or disciplinary-specific) considerations, or the need to nurture more data savvy habits and norms; many discuss the role of modeling drivers while others highlight supportive design elements. Put together, this variation provides the reader a rather comprehensive appreciation of the multidimensional pedagogical potential of data modeling centered pedagogies, that can be implemented to attend to the variety of educational needs characterizing the big data era. Though extensive, the account of this multidimensional phenomenon is not an exhaustive one, and we urge readers to further extend and build on the research depicted in the following pages.

Acknowledgements

Many colleagues have greatly supported the process of editing this collection of chapters. We are particularly grateful to Tim Erickson who has played a key role in improving the quality, the formatting, and the language of the chapters of the non-native authors with his very helpful, constructive, and intensive feedback. We thank all the reviewers for their constructive and helpful reviews to improve the quality and writing of the chapters within this book. We are very grateful to all authors and the Minerva School participants who contributed to this book—without them, this book would not have been possible. Finally, we express our sincere gratitude to the Minerva Foundation for their generous support in

funding the Minerva School 2020 and this book. Their commitment to advancing education has played a vital role in making this conference a reality. Their dedication to promoting knowledge exchange and fostering collaboration within the academic community is truly commendable.

References

Ben-Zvi, D., Gravemeijer, K., & Ainley, J. (2018). Design of statistics learning environments. In D. Ben-Zvi, K. Makar. & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 473–502). Springer.

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. A. Clements, A. J. Bishop, C. Keitel-Kreidt, J. Kilpatrick, & F. K.-S. Leung (Eds.), *Third international handbook of mathematics education* (pp. 643–689). New York: Springer Science + Business Media.

Budgett, S., & Pfannkuch, M. (2018). Modeling and linking the Poisson and exponential distributions. *ZDM Mathematics Education, 50*(7), 1281–1294. https://doi.org/10.1007/s11858-018-0957-x

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1).

Dvir, M., & Ben-Zvi, D. (2018). The role of model comparison in young learners' reasoning with statistical models and modeling. *ZDM Mathematics Education, 50*(7), 1183–1196.

Dvir, M., & Ben-Zvi, D. (2023). Informal statistical models and modeling. *Mathematical Thinking and Learning, 25*(1), 79–99.

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal, 16*(1), 44–49.

Erickson, T., Wilkerson, M., Finzer, W., & Reichsman, F. (2019). Data moves. *Technology Innovations in Statistics Education, 12*(1).

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education, 7*(2).

Garfield, J., & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice*. New York City, New York: Springer.

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal, 16*(1), 22–25.

Gravemeijer, K. (1999). How emergent models may foster the constitution of formal mathematics. *Mathematical Thinking and Learning, 1*(2), 155–177.

Hesse, M. B. (1962). *Forces and fields: The concept of action at a distance in the history of physics*. Mineola, NY: Dover.

Hancock, C., Kaput, J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27*, 337–364.

Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 193–215). Reston, VA: National Council of Teachers of Mathematics.

Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction, 14*, 69–108.

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (Vol. 5, pp. 101-159). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Lesh, R., & Doerr, H. M. (Eds.), (2003). *Beyond constructivism: Models and modeling perspectives on mathematics teaching, learning, and problem solving*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lesh, R., Hoover, M., & Kelly, A. (1992). Equity, assessment, and thinking mathematically: principles for the design of model-eliciting activities. In *Developments in school mathematics around the world* (Vol 3), Proceedings of the Third UCSMP International Conference on Mathematics Education (pp. 104–129). Reston: NCTM.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*(1–2), 152–173.

Manor, H., & Ben-Zvi, D. (2015). Students' emergent articulations of models and modeling in making informal statistical inferences. In *Proceedings of the Ninth International Research Forum on Statistical Reasoning, Thinking and Literacy* (pp. 107–117). Paderborn, Germany: University of Paderborn.

National Research Council (2012). A framework for K–12 science education: practices, crosscutting concepts, and core ideas. Washington, DC: The National Academies Press. https://doi.org/10.17226/13165.

Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modeling to connect data, chance and context. *ZDM - International Journal on Mathematics Education, 50*(7), 1113–1123. https://doi.org/10.1007/s11858-018-0989-2

Rubin, A., & Hammerman, J. K. (2006). Understanding data through new software representations. In G. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance: 68th NCTM Yearbook* (pp. 241–256). Reston, VA.: National Council of Teachers of Mathematics.

Sainani, K. L. (2014). Explanatory versus predictive modeling. *PM&R, 6*(9), 841–844.

Stillman, G., Kaiser, G., Blum, W., & Brown, J. (Eds.). (2013). *Teaching mathematical modeling: Connecting research to practice*. Dordrecht: Springer.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion papers). *International Statistical Review, 67*(3), 223–265.

Young learners' perspectives on the concept of data as a model: what are data and what are they used for?

Susanne Podworny and Daniel Frischemeier

Paderborn University, University of Münster
podworny@math.upb.de, dfrische@uni-muenster.de

Addressing data as a model and promoting a deeper understanding can provide a solid foundation for future modeling activities and enhance comprehension of data-based AI methods. Data—viewed not just as raw information but as a model—simplify complex aspects of the world and serve various purposes. A series of lessons taught 6th-grade students the AI method of decision trees using nutritional data. After the series of lessons, students answered a questionnaire and offered their perspective on data. The answers were analyzed using qualitative content analysis. Several students recognized data as a representation with context and understood its purpose for decision-making. However, some perspectives were incomplete, indicating a need for more explicit discussions about data as a model in the classroom.

Introduction

Our everyday life requires constant engagement with models based on data as indicated by current trends in statistics education (Burrill & Pfannkuch, 2023). Diagrams, studies, AI algorithms, predictive models, all of these can be found in our daily lives and can shape the statistical thinking and reasoning of young learners. The bases for all of these are data. However, data do not simply exist, but rather are themselves already models of reality (Konold et al., 2017). Good data is an important prerequisite for good models that improve our daily lives instead of, for example, fomenting prejudice (see, e.g., nasty algorithm as in Zweig, 2022). Data are being accumulated en masse nowadays, for example when using smartphones, while in the household, or in traffic. In addition, data are collected or generated specifically to pursue certain interests, such as being able to place targeted advertisements in a news feed. AI systems operate in the background for this purpose, recognizing patterns in data and making predictions or suggesting decisions. To gain some understanding of these processes, there has been a recommendation to teach Data Science and AI in school (Biehler & Schulte, 2018; Ridgway, 2016). To build appropriate understanding about how such AI systems work, students must see into the "engine room" and behind the black box of AI and the accompanying data exploitation.

The recently published German Data Literacy Charter (Schüller et al., 2021) emphasizes the importance of bringing the topics of data literacy and data-based decision-making into schools:

> "In concrete terms, this requires the inclusion of data literacy in the curricular and educational standards of schools, teacher training, and higher education. Learners should not only be addressed as passive consumers of data. We rather enable them to actively shape data-related insights and decision-making." (Schüller et al., 2021, p. 3)

For schools, appropriately simplified subject content and suitable examples are necessary that can be used to foster data literacy. A prerequisite for data literacy is the concept of data and understanding its nature. Learning concepts is more than learning a definition; there is a need to perform operations and observations in a learning situation (Gagné, 1965). So learning about the concept of data includes coping with, using, analyzing, and recognizing data.

Nutritional information in form of nutritional variables like calories, fat, sugar, etc. and the corresponding values for a food item can be seen as a model for food. These data models for several food products can serve as examples of an accessible use of data to support young students to develop a procedure (itself a model) that, similarly to formal AI algorithms, makes decisions about whether a food should be recommended or not. This served as the inspiration for a series of lessons in the ProDaBi project (www.prodabi.de/en) that was developed and implemented in several classes, to teach 6th grade students to use the AI method of decision trees with nutritional data (Podworny et al., 2021). As data are the basis for AI models and students are situated in a nutritional context in this case, students' notions of the concept of data are important and it is important what can and cannot be done with data.

Background

Data

"Data" is the first fundamental idea in statistics as described by Burrill & Biehler (2011). This includes aspects such as types of data, ways of collecting data, and measurement. More than 20 years ago, Cobb & Moore, (1997, p. 801) defined "*data are not just numbers, they are numbers with a context*" (emphasis in original) and this is still true today. Cobb & Moore give an example, a sequence of numbers (3, 5, 23, 37, 6, 8, 20, 22, 1, 3), which in itself is not very informative, and it is difficult to ascribe meaning to it. However, knowing that these are monthly numbers of people accused of witchcraft in Essex County, Massachusetts, from February 1692 onwards, reveals two waves of witch hunts in the USA colonial period. When we know the context of the witch hunts, the non-informative number sequence tells a story. This understanding about data is a prerequisite for modeling the world through data, for example, by displaying the data in a diagram like the one in Figure 1. The caption gives the diagram a meaning beyond the pure numbers.

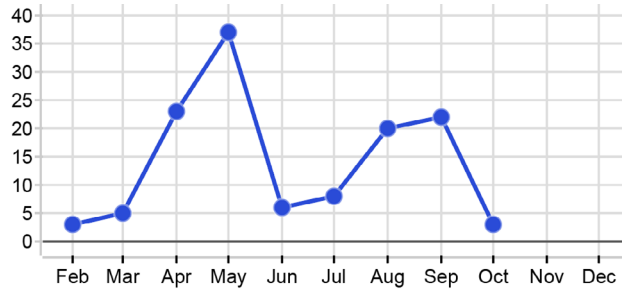


Figure 1. Number of people accused of witchcraft in 1692 in one village

Thus the context of any data is particularly relevant for getting information from the data, and integrating statistical and contextual information is a crucial part of statistical reasoning:

> "Context determines how and what data to collect, as well as how to analyze the data and interpret the results. This results in a constant interplay between considering a statistical problem and the context of the problem" (Weiland, 2019, p. 19).

This is stated by Wild & Pfannkuch (1999, p. 225) as well: "the ultimate goal of statistical investigation is *learning* in the *context* sphere" (emphasis in original). This view of what data are facilitates connecting what data are with an additional central entity in statistics: models.

Data as model

A model is used "to represent aspects of the world for various purposes" (Giere, 2004, p. 747). Applying this definition to data, we see that data essentially contain two features. First, data represent some aspects of the world, including the context of the data (Weiland, 2019; Wild & Pfannkuch, 1999); and second, data serve a purpose (Ainley, 2012). One way for using a model to represent aspects of the world is to exploit similarities between the model and the aspect of the world it represents. As a result, data fulfill two important characteristics of models: they are simplified representations of a more complex world (i.e., the context), and they serve various purposes as indicated in the definitions of a model by Ainley (2012) and Giere (2004).

To emphasize how data are also models, one is tempted to use a phrase such as "data models"—which would be a pleonasm—in order to emphasize that role. That would cause some problems here, however, for two reasons: first, other authors (e.g., Gafny & Ben-Zvi 2024, **in this volume on page 69**) use that phrase to mean other things. Second, as we will see shortly, there is another layer of modeling essential to the discussion, and using the word "model" to represent both layers would be confusing. We therefore ask the reader to remember throughout this chapter that we are investigating student conceptions of the model-nature of data—both context and purpose.

Models based on data: the next layer

While data values by themselves offer one representation of a phenomenon, the detailed nature of data and the variation that characterize it might render it not useful enough for the purpose for which the data were collected. To extract additional meaning, additional models that are more summative or aggregate in nature are often necessary. Thus, we use the data (as model) as the basis for these additional models, so that they become models of models. With the help of statistical thinking and reasoning (Burrill & Biehler, 2011), sometimes more information can be extracted from data when additional models are created based on it.

For example, for a food item modeled by all its nutritional data, it is hard to decide whether it should be "recommended" or not by just looking at individual values. A decision tree, as one model of machine learning based on data, can be prepared to jointly consider the data. Here, for example, the method of machine learning decision trees can help to create a model based on the data models and make predictions whether a new food item is more recommendable or not. Looking at the field of AI and machine learning, all models created there are data-driven (Hastie et al., 2009). Therefore, data play a fundamental

role in this process. A decision tree—a model of a decision algorithm—is a hierarchical structure created from data (Breiman et al., 1998). The data, in turn, are models for individual cases of food items. If the tree structure is not too large, then it is a well-suited model for teaching (Martignon et al., 2003) because it is transparent and easy to interpret. Still, the data are foundational, so there is a need for students to acknowledge that the whole process is based on the data.

Young students' perceptions of data as model

How young learners perceive data and their understanding of the concept of data is a topic that is not explicitly addressed in most research (English, 2014). It is much more often about how students interact with data in various aspects. For example, there are studies on the challenges young learners face in representing data (Harradine & Konold, 2006) or how they intuitively organize data (Konold et al., 2017). There is also research on what young learners use data for in different learning environments, e.g., for modeling a randomization test (Biehler et al., 2015) or making decisions (Engel et al., 2018) or how they understand concepts of data-based machine learning (Hitron et al., 2019). Viewing data as models (Lehrer & Schauble, 2007) is an underlying concept in most of this research but seldom a research topic in itself. Because conceptual learning is essential and involves the ability to construct commonalities and differences in order to build structured knowledge (Zeithamova et al., 2019), it is worth examining the student's concept of data.

Research questions

This study examines students' perspectives on data (as models) and models based on data after they engaged in a teaching sequence about data modeling using the example of decision trees.

Therefore, in this chapter, we pose the research question, "What are young learners' perceptions of data?" As stated above, data are models that have two characteristics: they represent the world by simplification and they have one or more specific purposes, e.g., to offer a classification prediction. To answer the research question, we explore these two aspects in two sub-questions.

1. What ideas do young learners express about *what data are* after a teaching unit on decision trees on food data?
2. What ideas do young learners express about *what data are used for* after a teaching unit on decision trees on food data?

Method

The teaching sequence and its implementation

To provide young learners (in this study, grade 6, ages 11–12) with a meaningful engagement with data modeling that can potentially allow them to experience data as models, we designed a series of lessons on creating, applying, and using decision trees. The principal question of the series of lessons was: How can we use nutrition information to predict food as "recommended" or not? Data of food items were chosen as the basis for a model that combines that information to predict "recommendation." Data-based decision trees were built manually in class for the prediction to develop the notion that a decision tree is a model based on data.

The foundation for the lesson is a set of data cards like the one in Figure 2, left, that we regard as a model of a *case*, which here is a food product (an apple in this instance). Several of these cases together can be the data that form the basis for deciding whether a food item is recommended or not. A decision model based on these data can be a tree that predicts recommendations for new food items, such as the two-level decision tree in Figure 2, right.

For implementation in class, approximately 30 of these cards were labeled by the students with green or red paper clips to indicate if the item was recommended or not as the first modeling activity as suggested by Hitron et al. (2019). Students then used the dataset to create two-level decision tree models to decide, for new foods, whether they were more likely to be recommended. By the end, students were able to create decision tree models like the one in Figure 2 (right) using data, test them, and apply them to new data. In total, the lesson series included eight lessons of 45 minutes each. Several cycles of the teaching series have taken place since May 2021.

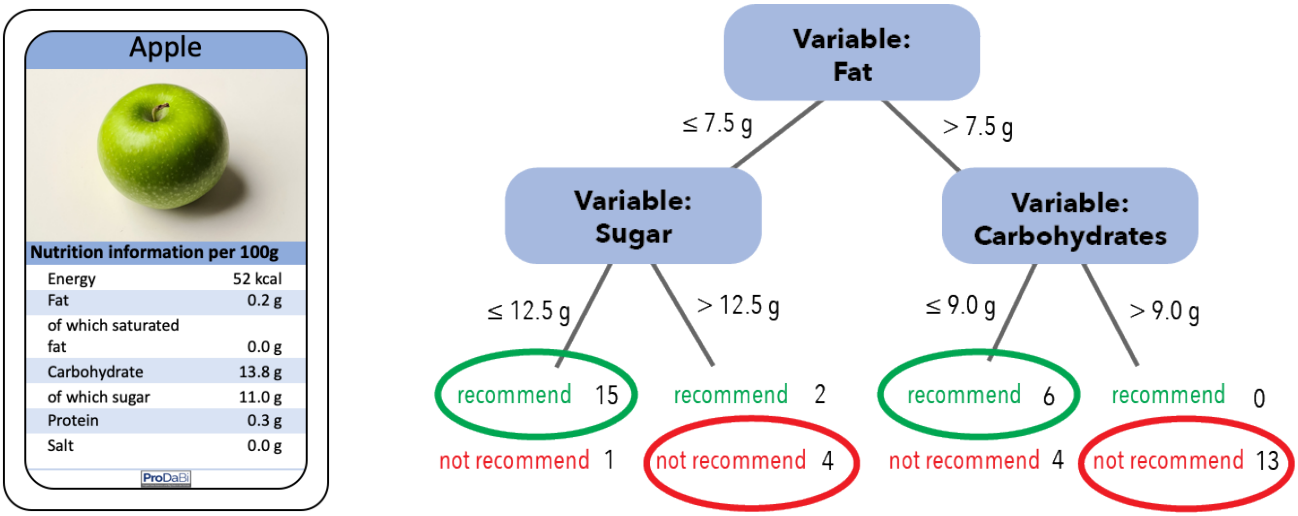


Figure 2. Example for a data card (left) and a decision tree model based on food data (right)

Study design and data collection

The data that model food items, from which the decision trees were created, were a fundamental element of the lesson series. However, the concept of data-as-model was not explicitly discussed. Therefore, it was necessary to investigate what perspectives the students revealed on data after the series of lessons.

As part of the research on the series of lessons, a questionnaire was developed for the students and administered after the last lesson. The questionnaire included a variety of questions, one of which is the focus of this chapter: "What are data and what do you need data for?" 208 of 263 learners who answered the questionnaire gave an answer towards this question. The students' answers were evaluated as part of the research report in the present chapter. The questionnaire was implemented online on a server of the University of Paderborn and all answers of the students were available digitally.

Participants

Since the start of the series of lessons in May 2021, 263 students have completed the questionnaire, of which 149 are female, 113 are male and one has no indication. The participants were all sixth-grade students, ages 11–12, from different secondary schools in Germany and all of them had participated in the teaching series. The students had no previous knowledge of computer science or statistics before the series of lessons; one class had previously studied the topic of nutrition in biology lessons.

Data and methods of data analysis

In order to find out what the learners' ideas about data were, the answers to the question "What are data and what do you need data for?" were evaluated using qualitative content analysis (Mayring, 2015). The aim of this systematic and rule-guided evaluation was to identify structures in the answers and to draw conclusions about the learners' ideas about data. The statements were coded independently by the two authors of this paper. In case of unequal coding, discussions were held until agreement was reached.

The responses were coded in two steps. In the first step, we coded aspects of the written answers that addressed the questions *what* are data (code 1), and aspects that addressed the question *what* do you need data *for* (code 2). Answers that addressed neither were also coded (code 3). In the second step, the partial answers to the two questions were analyzed individually.

First, we looked at all (partial) sentences that we coded in the first step as "what," indicating students' view of data as a representation. We deductively defined the codes 1-1 and 1-3 and added 1-2 and 1-4 inductively. See Table 1.

In the same way we looked at everything we coded in the first step as "what for" to categorize learners' answers for "What do you need data for?" This showed students' perspectives on the purpose of data. We deductively defined codes 2-1 and 2-2 in Table 2. Inductively, we added codes 2-3, 2-4 and 2-5 in Table 2.

Code	Definition	Example
1-1 Data as representation by numbers with context	Learners describe data as numbers with context or give specific examples of numbers with context (Cobb & Moore, 1997).	"Data are for example how much fat a strawberry contains." (student 231)
1-2 Data as representation by numbers	Learners describe data as numbers only.	"Data are numerical values" (student 68)
1-3 Data as representation by statistical characteristics	Learners describe data using statistical terms but more than just "numbers."	"For example, in the case of a food product, data are the attributes and the values." (student 35)
1-4 Data as information	Learners generally describe data as information without further explanation.	"Data are important information" (student 85)

Table 1. Coding manual for "What are data?"

Code	Definition	Example
2-1 Purpose: specific information	Learners describe the use of data to have (more) information about specific aspects of the world, e.g. about a person, a food item, etc. (Wild & Pfannkuch, 1999)	"You need data to know how much of the attribute the food has in it." (student 131)
2-2 Purpose: Decision	Learners describe the use of data as the basis for a decision, classification, or an artificial intelligence. (Breiman et al., 1998)	"You need the data to decide whether the food is healthy or unhealthy." (student 234)
2-3 Purpose: general information	Learners describe the abstract use of data to have (more) general information.	"With the help of data, you can figure things out." (student 108)
2-4 Computer need data	Learners describe the use of data with exclusive reference to the computer.	"Data is needed, for example, to save things [in the computer]." (student 159)
2-5 Other	Learners describe the use of data in none of the other categories.	"You need data, because you can't do without them (I think)." (student 140)

Table 2. Coding manual for "What do you need data for?"

Results

Out of the 208 learners who answered the question, 51 made a statement only concerning the aspect of "what" are data, 88 only for "what for" and 69 learners answered both. In sum there were 120 statements for "what" and 157 for "what for." Those were analyzed separately according to the manual. We introduce the detailed findings in the next sections.

Results for Subquestion 1

The categorization of answers for the question "What are data?" is shown in Figure 3.

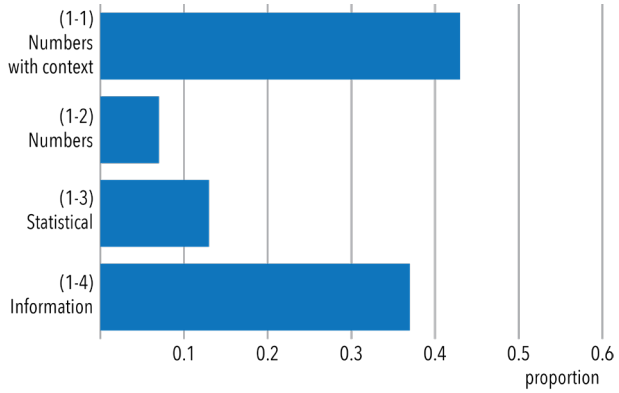


Figure 3. Proportion of students (out of N = 120) in each code for "What is data?"

Answers that could be assigned to the category "data as representation by numbers with context" (1-1) occurred most frequently. A total of 51 students (43% of 120) described data in this way. The responses often included a reference to the food topic of the lesson series. Here it could be interpreted that the students attributed a certain model character to data. In addition, there were another eight students (7% of 120) who described data only as "data as representation by numbers" without any reference. These descriptions were possibly influenced by the series of lessons, which only involved numerical data.

A different perspective was taken by students who described data in terms of their statistical properties (code 1-3), such as the terms "variables," "attributes," or in circular reasoning as "data are data." These responses did not attribute a model nature to the data. Finally, there was a large group of students (37% out of 120) who described data abstractly as "information." One explanation for these expressions might lie in the fact that the teaching occurred in a computer science lesson, and this perspective may originate from the computer science standpoint. At least in Germany, "data and information" are often mentioned in the same breath in computer science education. It is unclear again if these responses reflect attributing a model nature to the data or not.

All statements of learners regarding what data are operated on different levels. Some were extremely general, so that it was difficult to confidently identify an underlying perception; others were specifically related to the content of the lesson series and mentioned the variables and values dealt with there as representations. Neither the word model nor representation appeared in any answer. The majority of the learners understood data in the sense of Cobb & Moore (1997), as information or belonging to a certain context.

Results for Subquestion 2

Figure 4 reflects the results of the evaluation of the answers to "What do you need data for?" to identify students' perspective on the purpose of data.

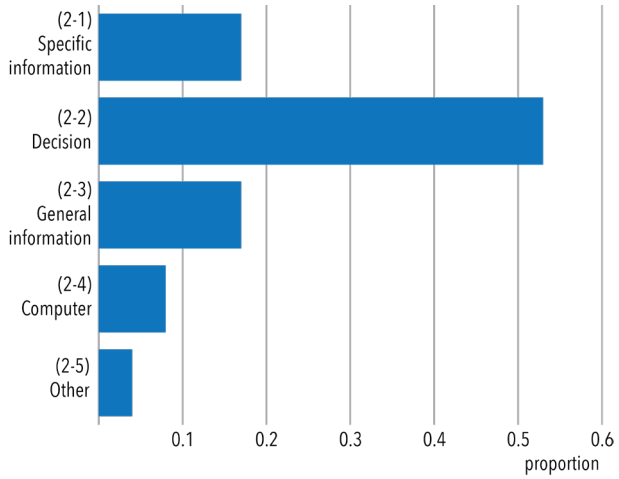


Figure 4. Proportion of students (out of N = 157) in each code for "What do you need data for?"

Prominent in Figure 4 is that 53% of the students who gave an answer to this part of the question reported the decision-making nature as the purpose for data. This subsumed responses that mentioned "deciding," "classifying," or creating an AI with these goals in mind. This perspective on the use of data seemed to have been inspired by the series of lessons where the goal was to create decision trees as a method of AI based on data. It can be interpreted that the perspective prevailing here was that data's purpose is to create an AI model.

The responses categorized in (2-1) took the use of data as a model into consideration, so they fit more with Cobb and Moore's (1997) view and added the purpose of data in contrast to category (1-1).

Again, there were students who took a very general perspective and generally described the use of data as "to figure something out" without specific reference to any context (code 2-3, 17% of 157). However, meaningful figuring out only exists with a purpose, so that here too, with caution, a model perspective could be interpreted from the answers.

Students who saw the use of data in the fact that something can be "stored digitally" (student 218) or that "an application can be programmed" (student 26) (code 2-4) seemed to see a high connection between data and computers or a digital representation. This might be a more limited view, neglecting any context or purpose.

The answers that were categorized as "other" (2-5) were a low proportion of all answers with 4% of the total. A relationship to data models could hardly be ascribed to these answers.

Combined results concerning the research question

The overall research question was "What are young learners' perceptions of data?" Considering the perspective of data models that have a representational characteristic (subquestion 1) and a purpose (subquestion 2), there was a need to look at both characteristics at once in the answers.

Of the students, 38 (18% of 208 students who gave any response at all) gave a response that described both the representational role and the purpose of data. This included all answers that were categorized as code 1-1, 1-2 or 1-3 and with 2-1, 2-2, 2-3, i.e., all answers that in some way described a representational character and a purpose character to data. This left a different 38 students (18% of 208) who only considered the representational nature of data and 95 students (46% of 208) who only addressed the purpose of data. All other students were categorized either as 1-4 (data as information) or 2-4 (data are needed for the computer) or 2-5 (other).

Discussion

Data are available nearly everywhere and used for nearly everything, so there is a need to critically reflect on what data are as is explained in detail by Engel (2024, **in this volume on page 101**).

Even despite the students' experience during the lesson series, many of them still expressed naïve views on data. Students have a notion of what data are and what they are used for. In the learning sequence that the students completed, data were used constantly to create decision trees. The perspective of data as models in the sense defined above was not explicitly addressed, as is probably the case in many learning sequences. This is one main motivation to investigate what perspectives on the concept of data these students had and what purpose students attributed to data at the end of an intervention that meaningfully, however implicitly, engaged them with data models (Stillman & Brown, 2023).

The model aspect was not explicitly addressed in class, but subliminally plays an important role in the students' ideas of what data are and what they are needed for. In fact, a large proportion of the students made statements that implicitly addressed some aspect of data as models. However, the analyses show that more than 80% of the students, although being frequently exposed to data in their everyday lives, have a rather incomplete picture of data as models. Their concept of data as models is quite oversimplified, as one might expect for the beginning of conceptual learning (Feldman, 2003).

It is also important to consider the 55 students who did not answer the question at all. This is 21% of all 263 students who either would not or could not give an answer. Since all the students had completed other answers in the questionnaire, it could be concluded that this fifth of the students chose not to answer this question. This and the other findings advocate supplementing the students' experience with more explicit discussions about the nature of data. In order for this perspective by Cobb and Moore (1997) and also basic perspective by Wild and Pfannkuch (1999) to be taken for modeling in statistics, it should be more explicitly addressed in the classroom. In this way, incomplete or inconsistent views can also be countered, as they also emerged in the analysis.

The implication is that it may be useful to talk explicitly about the fact that data are already models in every subject where data are used. Data are representations that represent certain aspects of the world, but not all, and have a purpose. The use of data as described by Giere (2004) is not predetermined by the data model, but by the person who uses data as a model for something specific (Stillman & Brown, 2023). And this should be communicated transparently to students to overcome their sometimes naïve view (Dvir & Ben-Zvi, 2021).

Data form the basis of all further models and a mature understanding of the nature of data is thus necessary for modeling activities, as was also investigated by Büscher (2024, **in this volume on page 49**). Gafny and Ben-Zvi (2024, **in this volume on page 69**) have designed a framework that can support young learners to gain a deeper understanding of data models and modeling. If a good foundation is laid in terms of students' understanding of the concept of data, it is likely that this can be better built upon later. Emphasizing the concept of data as models, and including statistical concepts such as signal and noise (Burrill & Biehler, 2011; Gould 2024, **in this volume on page 81**), would certainly also promote a more mature view of the concept of data and an understanding of data-based AI methods.

References

Ainley, J. (2012). Developing purposeful mathematical thinking: a curious tale of apple trees. *PNA*, *6*(3), 85–103.

Biehler, R., Frischemeier, D., & Podworny, S. (2015). Preservice teachers' reasoning about uncertainty in the context of randomization tests. In A. Zieffler & E. Fry (Eds.), *Reasoning about uncertainty: Learning and teaching informal inferential reasoning* (pp. 129–162). Catalyst Press.

Biehler, R., & Schulte, C. (2018). Perspectives for an interdesciplinary data science curriculum at German secondary schools. In R. Biehler, L. Budde, D. Frischemeier, B. Heinemann, S. Podworny, C. Schulte, & T. Wassong (Eds.), *Paderborn symposium on data science education at school level 2017: The collected extended abstracts* (pp. 2–14). Universitätsbibliothek Paderborn. https://doi.org/http://doi.org/10.17619/UNIPB/1-374

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1998). *Classification and regression trees*. Chapman & Hall.

Burrill, G., & Biehler, R. (2011). Fundamental Statistical Ideas in the School Curriculum and in Training Teachers. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics—Challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 57–69). Springer Science+Business Media. https://doi.org/10.1007/978-94-007-1131-0_10

Burrill, G., & Pfannkuch, M. (2023). Emerging trends in statistics education. *ZDM Mathematics Education*. https://doi.org/https://doi.org/10.1007/s11858-023-01501-7

Büscher, C. (2024). **"Design principles for developing statistical literacy by integrating data, models, and context in a digital learning environment" in this volume on page 49.**

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

Dvir, M., & Ben-Zvi, D. (2021). Students' actual purposes when engaging with a computerized simulation in the context of citizen science. *British Journal of Educational Technology*, *53*(5), 1202–1220.

Engel, J. (2024). **"Some reflections on the role of data and models in a changing information ecosystem" in this volume on page 101.**

Engel, J., Erickson, T., & Martignon, L. (2018). Teaching and learning about tree-based methods for exploratory data analysis. Looking back, looking forward. *Proceedings of the Tenth International Conference on Teaching Statistics* (ICOTS10, July, 2018), Kyoto, Japan.

English, L. D. (2014). Establishing statistical foundations early: Data modeling with young learners. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA* (pp. 1–6). International Statistical Institute.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, *12*(6), 227–232. https://doi.org/https://doi.org/10.1046/j.0963-7214.2003.01267.x

Gafny, R. & Ben-Zvi, D. (2024). **"Reimagining data education: Bridging between classical statistics and data science" in this volume on page 69.**

Gagné, R. M. (1965). The learning of concepts. *The School Review*, *73*(3), 187–196. https://www.jstor.org/stable/1083668

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, *71*(5), 742–752. https://www.jstor.org/stable/10.1086/425063

Gould, R. (2024). **"Traditional statistical models in a sea of data: teaching introductory data science" in this volume on page 81.**

Harradine, A., & Konold, C. (2006). How representational medium affects the data displays students make. International Conference on Teaching Statistics-7, Brazil.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer. https://doi.org/https://doi.org/10.1007/978-0-387-84858-7

Hitron, T., Orlev, Y., Wald, I., Shamir, A., Erel, H., & Zuckerman, O. (2019). Can children understand machine learning concepts?: The effect of uncovering black boxes. *Prodceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/https://doi.org/10.1145/3290605.3300645

Konold, C., Finzer, W., & Kreetong, K. (2017). Modeling as a core component of structuring data. *Statistics Education Research Journal*, *16*(2), 191–212.

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 149–176). Lawrence Erlbaum Associates.

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. R. (2003). Naive and yet enlightened: from natural frequencies to fast and frugal decision trees. In D. Hardman & L. Macchi (Eds.), *Psychological perspectives on reasoning, judgment and decision making* (pp. 189–211). John Wiley & Sons. https://doi.org/https://doi.org/10.1002/047001332X.ch10

Mayring, P. (2015). Qualitative content analysis: theoretical background and procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education* (pp. 365–380). Springer.

Podworny, S., Fleischer, Y., Hüsing, S., Biehler, R., Frischemeier, D., Höper, L., & Schulte, C. (2021). Using data cards for teaching data based decision trees in middle school. 21st Koli Calling International Conference on Computing Education Research (Koli Calling '21), November 18–21, 2021, Joensuu, Finland.

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

Schüller, K., Koch, H., & Rampeld, F. (2021). Data literacy charter. Stifterverband. https://www.stifterverband.org/sites/default/files/data-literacy-charter.pdf

Stillman, G. A., & Brown, J. P. (2023). Modeling the phenomenon versus modeling the data set. *Mathematical Thinking and Learning*, *25*(3), 270–295. https://doi.org/10.1080/10986065.2021.2013144

Weiland, T. (2019). The contextualized situations constructed for the use of statistics by school mathematics textbooks. *Statistics Education Research Journal*, *18*(2), 18–38.

Wild, C., & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, *67*(3), 223–265.

Zeithamova, D., Mack, M. L., Braunlich, K., Davis, T., Seger, C. A., van Kesteren, M. T. R., & Wutz, A. (2019). Brain mechanisms of concept learning. *The Journal of Neuroscience*, *39*(42), 8259–8266.

Zweig, K. (2022). *Awkward intelligence: Where AI goes wrong, why it matters, and what we can do about it*. MIT Press.

Modeling situations with two binary events and different visualizations

Karin Binder

LMU Munich, Theresienstraße 39, 80333 Munich
Karin.Binder@lmu.de

Situations with two binary events can be modeled with the help of different graphical models, like tree diagrams, double trees, 2×2 tables, or net diagrams. Previous studies have shown that not all types of graphical models are equally helpful to learners. Furthermore, with tree diagrams and double trees, the conditional probabilities tend to dominate, while with 2×2 tables, joint probabilities come into focus. This chapter presents a relatively new graphical model that shows conditional probabilities and joint probabilities simultaneously: the net diagram. We will then see empirical results from two different studies that investigated—with the help of paper-and-pencil-tests in a quasi-experimental design—whether the net diagram, as a comprehensive graphical model, overloads participants in solving tasks regarding conditional probabilities, and how typical errors in so-called Bayesian tasks depend on 1) the information format, and 2) the visualization used. Although the net diagram shows a large amount of information, it does not seem to overload learners cognitively any more than a frequency double tree. Furthermore, the typical errors in Bayesian tasks depend not only on the information format (natural frequencies vs. probabilities, e.g., "80 out of 100 women are ill" compared to "the probability of a woman being ill is 80%"), but also strongly on the visualization used. In summary, reasoning with models in Bayesian situations depends on the graphical model chosen. Knowing which errors are common "traps" could help students engaged in modeling distinguish correct from incorrect solutions. Based on the findings that joint probabilities, conditional probabilities, and inverted conditional probabilities are often confused with each other, the chapter concludes with an idea on how this student difficulty could be addressed using data cards.

Introduction

Gal (2024) emphasized that "people have to understand models and the results of modeling when reading or watching the news, broadly viewed, including newspapers and print media, websites of news organizations, Facebook, blogs, etc." (**in this volume on page 91**). Unfortunately, many people in society today are not yet able to understand such models, as shown by numerous examples in the COVID pandemic (Martignon et al., 2023). The pandemic clearly demonstrated the importance of modeling situations with two dichotomous characteristics, e.g., rapid test results (positive vs. negative) and the status of potential COVID infection (infected vs. uninfected; see also Martignon et al., 2023). Incorrectly interpreting statistical information in the real world can lead to overdiagnosis or overtreatment in medicine (Wegwarth & Gigerenzer 2013), or even suicide in the worst case, if too much trust is placed in a positive test result indicating a serious illness (Stine, 1996). In the field of law, incorrect modeling of statistical information even sometimes leads to false convictions (Fenton, 2011). Particularly difficult in this context are so-called "Bayesian situations," in which conditional probabilities have to be computed. These have been intensively researched, especially in cognitive psychology (e.g., Gigerenzer & Hoffrage, 1995; McDowell & Jacobs, 2017).

However, fortunately, there are strategies in Bayesian situations to help understanding the statistical information. In the following, two studies are presented that focus (1) on different graphical models for situations with two dichotomous characteristics (including one relatively new graphical model), and (2) the effect of presenting frequencies instead of probabilities. Here, the data to be modeled is already available in aggregated form, and these aggregate values must be combined to form conditional probabilities. Therefore, both studies focused on *graphical models* for the visualization of conditional probabilities. Finally, a more data-driven approach will be discussed that could improve Bayesian reasoning beyond graphical models.

Theoretical Background

Table 1 shows a typical Bayesian situation in the format of probabilities (left) and in the format of natural frequencies (right) for the context "COVID." These situations are represented in a later step with different graphical models that focus on different statistical information (e.g., conditional probabilities or joint probabilities, see Figures 2 and 3). Numerous empirical studies, and the meta-analysis from McDowell and Jacobs (2017), show the natural frequency effect, which means: Natural frequency versions of Bayesian reasoning problems (see right side of Table 1) are more often solved correctly than probability versions of the task (see left side of Table 1). While only about 5% of participants were able to solve the probability version of those tasks, about 25% of participants were able to solve the natural frequency version (McDowell & Jacobs, 2017). Therefore, learning environments that foster learners' reasoning skills in situations with two dichotomous events should refer to an imaginary sample and enable thinking in natural frequencies. Furthermore, there is evidence that the occurrence of typical wrong answers (e.g., in Table 1, the joint occurrence error: "480 out of 10,000 persons" instead of "480 out of 670 persons") also depend on the information format (i.e., probabilities vs. natural frequencies; Gigerenzer & Hoffrage, 2015; for an overview see Binder et al., 2020).

Bayesian tasks—like the COVID task in Table 1—can be understood as modeling tasks in the sense of Blum et al. (2007), as described by Eichler and Vogel (2015); their data modeling cycle appears as Figure 1. Referring to that

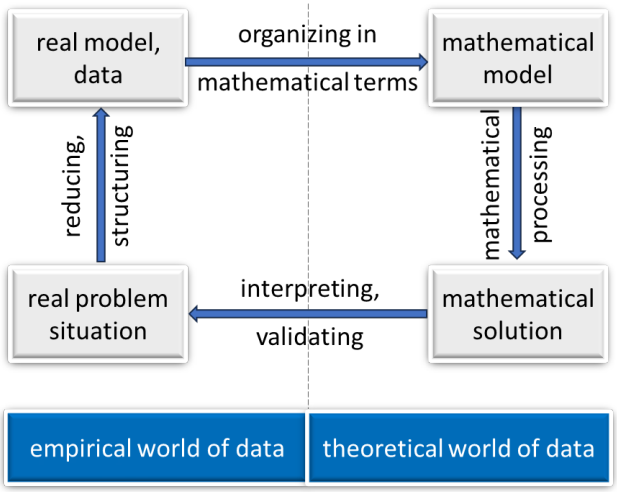


Figure 1: Data modeling cycle from Eichler & Vogel (2015)

figure, the real problem situation and the stated question must first be properly understood. In real-life situations, it is often even necessary to laboriously collect the information at first, and to be able to recognize whether individual pieces of information are still missing. The next step would be to extract the relevant information in the text (i.e., prevalence, sensitivity, specificity) and to build a real model by relating the given parameters to each other. Then the real-world information can be transformed into the mathematical world. In this translation process, graphical models can be used, such as 2×2 tables, tree diagrams, or double trees (Batanero & Sanchez, 2013; Khan et al. 2015; Böcherer-Linder & Eichler, 2019, Binder, Krauss & Wiesner 2020). Next, the model helps with obtaining mathematical results, such as conditional probabilities.

Probabilities	Natural frequencies
Statistics on persons who have likewise just returned from a high incidence area with symptoms of a cold (such a person is referred to as "a person" in the following) and then use the COVID self-test reveal:	
• There is a 5% probability that a person is infected with COVID. • If a person is infected with COVID, then the probability is 96% that this person tests positive. • If a person is not infected with COVID, then the probability is 2% that this person tests positive nevertheless.	• 500 out of 10,000 persons are infected with COVID. • Out of 500 persons that are infected with COVID, 480 receive a positive test result. • Out of 9,500 persons, that are not infected with COVID, 190 will nevertheless receive a positive test result.
Question: If a person tests positive, then what is the probability that the person is infected with COVID?	**Question:** How many persons test positive, and how many of those are actually infected with COVID?
Answer: 71.6%	**Answer:** 480 out of 670 persons

Table 1: Probability and natural frequency versions of a Bayesian reasoning task regarding COVID

Finally, the meaning of the mathematical result for the real world must be grasped. For example: What does the calculated value of 71.6% mean? Why is it so surprisingly low? What is the actual difference in the proportion of *infected persons among those who test positive* and the proportion of those who *test positive among those who are infected*? Which of the two pieces of information is more important for me?

In the following, I will focus on the mathematical model and situations, in which aggregate data is already available. Therefore, the given numbers and corresponding explanations in Table 1 can be seen as the real model, whereas the different visualization templates can be seen as the mathematical models. Depending on which graphical model is used for mathematical processing, certain aspects tend to come to the fore and other aspects tend to fade into the background. Numerous studies have shown that at least some types of visualization can promote Bayesian reasoning (Binder, Krauss & Bruckmaier, 2015; Pfannkuch & Budgett, 2017; McDowell & Jacobs, 2017), including in school teaching (see e.g., Wassner, Martignon & Biehler, 2004).

In teaching statistics at school and university, two visualizations are primarily used in modeling when situations with two dichotomous characteristics of a real model have to be converted into a mathematical model: 2×2

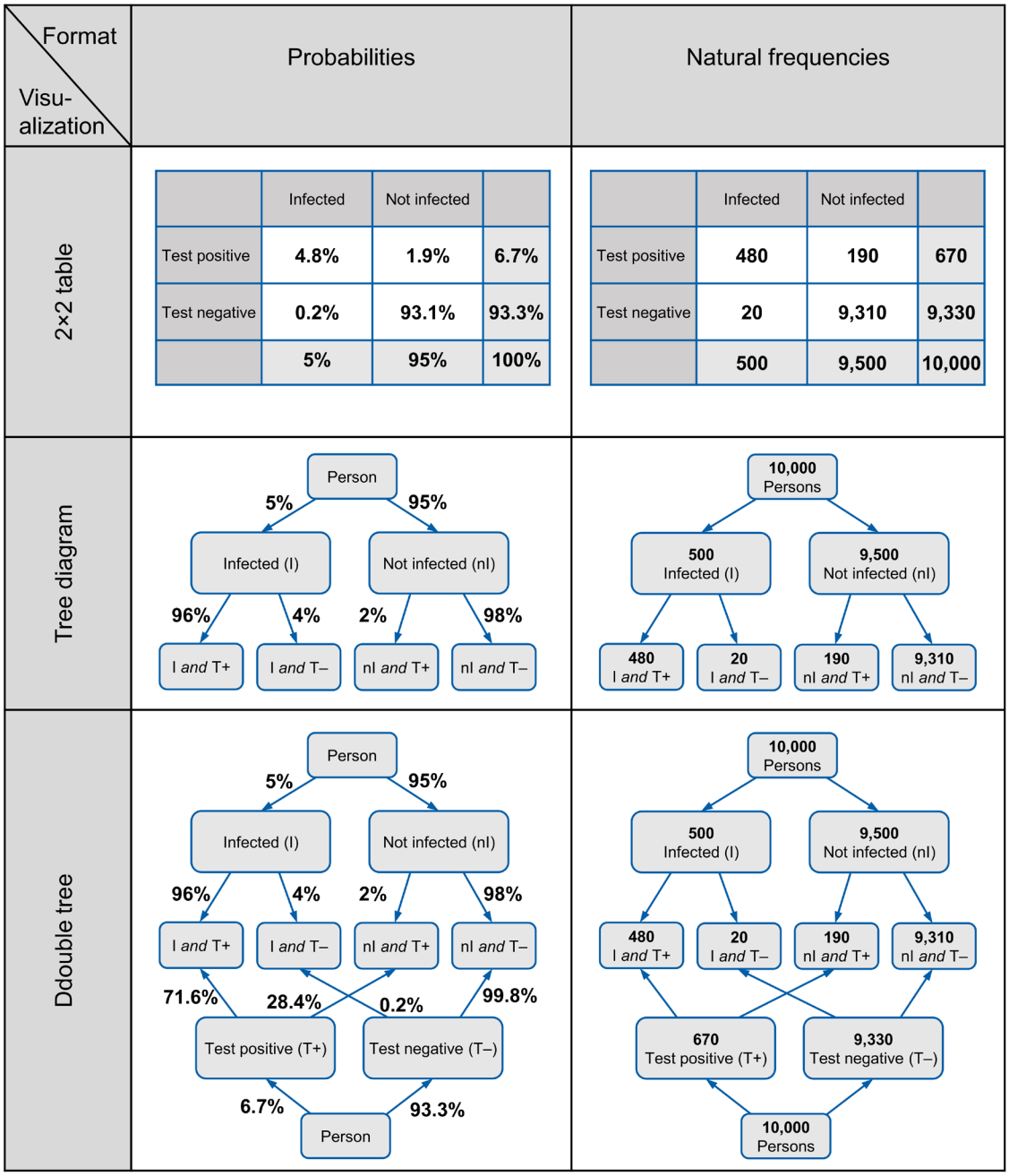


Figure 2: 2×2 table, tree diagram and double tree diagram as graphical model with probabilities (left) or natural frequencies (right) for the COVID problem.

tables and (double) tree diagrams. These visualizations can be depicted with probabilities or with frequencies. Figure 2 illustrates the COVID problem with the help of 2×2 tables, tree diagrams, and double tree diagrams in the probability version (left) and the natural frequency version (right).

Visualizations with frequencies have been shown to help students significantly better than probability visualizations (Binder, Krauss & Bruckmaier, 2015). That study compared the effectiveness of (1) natural frequencies and (2) visualization (e.g., tree diagrams, 2×2 tables) in helping people understand conditional probabilities in Bayesian tasks. 259 school students (11th grade) solved Bayesian tasks, which were presented either without any visualization, with 2×2 tables, or with tree diagrams. The data in each case were presented as either probabilities or natural frequencies. The result was that a maximum of 10% students were able to correctly solve a Bayesian task if the task was presented with probabilities, despite the fact that 2×2 tables with probabilities, and tree diagrams with probabilities at the branches, are a focus of stochastics teaching in school. Natural frequency trees (which are mostly unknown in German schools), on the other hand, were able to support students much better in solving the problem (correct solution rate 45%).

Because tree diagrams or double trees are node-branch structures, these two visualizations (compared to the 2×2 table) can even display probabilities on branches, and frequencies in nodes, simultaneously. This is an educational advantage, as it allows the frequency concept to be used to better understand probabilities. However, 2×2 tables and (double) trees each have a decisive

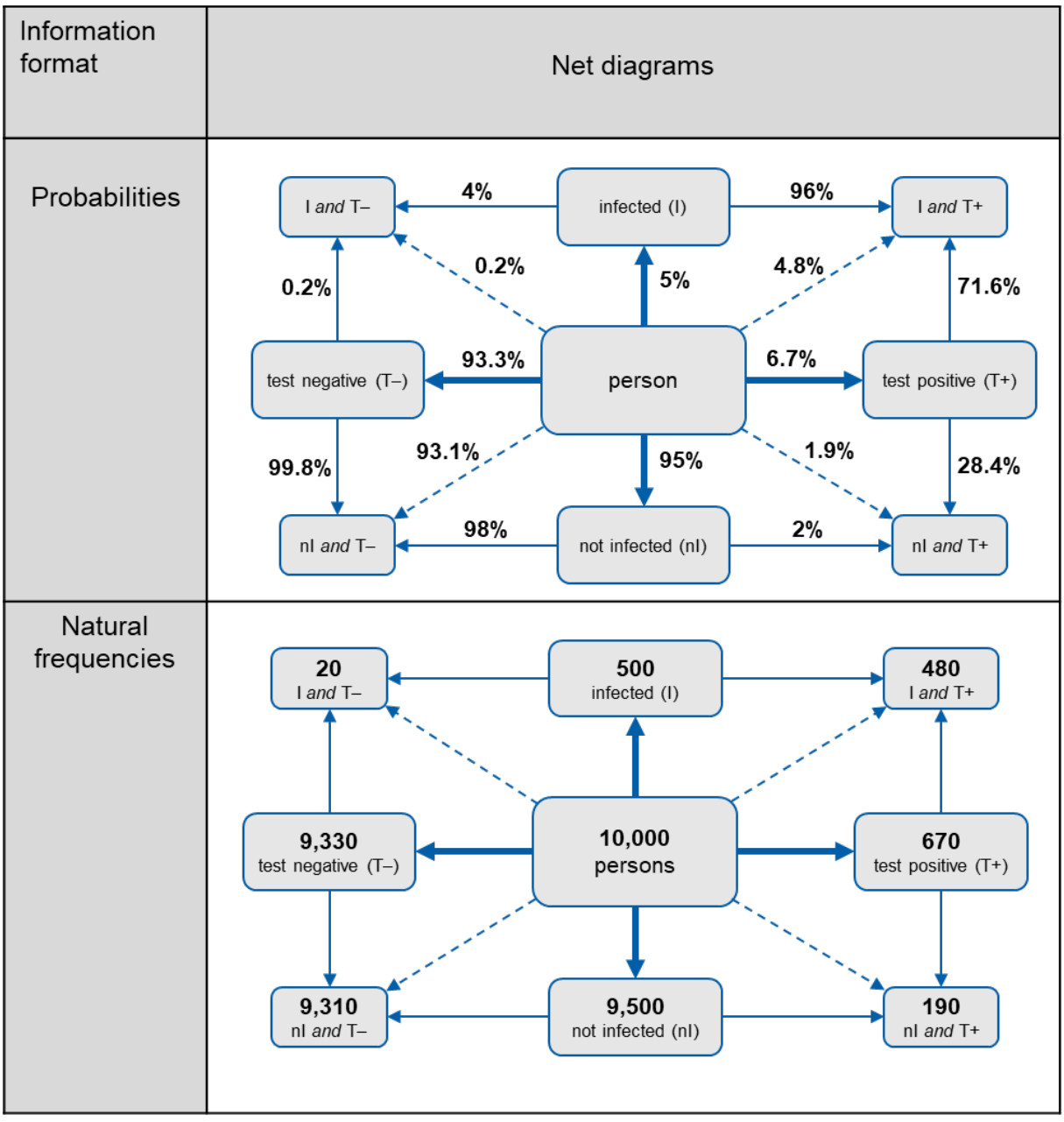


Figure 3: Net diagram as graphical model with probabilities (above), or frequencies (below).

disadvantage in the probability representation: In 2×2 tables, joint probabilities (e.g., $P(A \cap B)$) are represented, but no conditional probabilities (e.g., $P(A|B)$). It is exactly the other way around with (double) trees. The two—2×2 tables and (double) trees—are therefore particularly suitable for modeling only very specific starting situations.

Therefore, for modeling situations with two dichotomous events, a visualization is desirable which is equally suitable for the representation of joint probabilities and conditional probabilities. Figure 3 shows a relatively new visualization (also a node-branch structure), in which all absolute frequencies and all probabilities can be depicted at a glance: The frequency net. With the help of this visualization, all starting situations for two dichotomous events can be modeled. No matter which parameters are given in the real situation (marginal probabilities, joint probabilities, conditional probabilities)—they all can be represented within the frequency net in the modeling process during the transformation of a real model into a mathematical model.

In the following, two studies are described which examined the effectiveness of the frequency net as a graphical model in Bayesian reasoning situations in comparison with other visualizations. Study 1 (Binder, Steib & Krauss, 2022) focused on node-branch structures and investigated the question whether showing too much information negatively influences the solution process. Starting from a tree, proceeding to a double tree, and finally to a frequency net, successively more information is shown. On the one hand, this can help because more information can be read directly. On the other hand, the cognitive load may increase due to the increasing amount of information presented. Study 2 (Binder, Krauss & Wiesner, 2020) focused on typical errors and compared the frequency net with double trees and 2×2 tables.

Study 1–Too much information: Curse or blessing?

Considering the many options for graphical models in situations with two dichotomous features the question arises: Which of the many possible graphical models should be used in learning environments to model the data of the two features? As described above, the net diagram represents all relevant information in situations with two features: Four marginal probabilities, four joint probabilities and eight conditional probabilities. This multitude of information could be a blessing, but also a curse. The study outlined below explores this question.

The study (Binder, Steib & Krauss, 2022) focused only on node-branch structures—tree diagrams, double trees, and frequency nets—as graphical models for situations with two binary events. Each of these three visualizations can be seen as an extension of the previous visualization: A double tree involves all of the statistical information already presented in a tree diagram and supplements it with the inverted conditional probabilities and the missing marginal probabilities of the second event. Similarly, net diagrams involve all of the information represented in a double tree and supplements it with joint probabilities. This successive extension from a tree diagram to a double tree and then to a net diagram leads to a presentation of more and more information, which can have positive but also negative sides. Consider a typical Bayesian task in which a positive predictive value is calculated from three given parameters, as in Table 1. A probability tree displays the three given parameters (prevalence, sensitivity and false-positive rate); using those values, a student can calculate the positive predictive value (probability of being infected, if the test is positive) with the help of the addition rule and the multiplication rule. However, by extending the tree diagram to a double tree, the positive predictive value is already depicted within the double tree. So, there is more information provided—the very information we are interested in. By extending the double tree to a net diagram, joint probabilities are also shown. To show these four additional probabilities might be irritating and could confuse learners; especially since Study 2 will show that the confusion of $P(A \cap B)$ with $P(A|B)$ is one of the common mistakes in Bayesian reasoning problems.

Therefore, two different aspects should be distinguished: By extending the node-branch structure, the inference degree (i.e. the number of mental steps required) decreases. However, the complexity of the representation increases because additional, possibly irrelevant information is presented.

To explore this issue, this first study focused on the node-branch structures—tree diagram, double tree, and net diagram—first using natural frequencies, and afterwards using probabilities. With the successive extension of the typical tree diagram to the double tree and finally to the net diagram, the inference degree for questions about conditional probabilities *decreases* (i.e., fewer mental steps are required), however, at the same time the complexity of the representation *increases* and thus maybe the extrinsic cognitive load. The study examined which of these two effects predominates.

Method

In a paper-and-pencil-study (Binder, Steib & Krauss, 2022), 269 school students (grade 10) had to answer questions about conditional probabilities in typical Bayesian reasoning tasks. Each student received three questions, the first two in "natural-frequency format" (like the right side of Table 1), the third in "probability format" (like the left side). The questions were presented using text only, or using one of three visualizations—tree diagrams, double trees, or net diagrams. Any visualizations were already completely filled in with absolute frequencies or probabilities.

Results

When students used natural frequencies, the successive extension of the node-branch structures positively affected performance (see Figure 4A). Although double trees and nets were entirely unfamiliar to the students, these visualizations—which were already completely worked out—provided the best support to the students in completing the tasks.

In the probability format (see Figure 4B), participants performed best with the help of the completely filled double tree. However, the solution rate of 31% is low, especially considering that the correct solution actually appears in the visualization. Extending the double tree to a probability net reduced participants' performance to 23%.

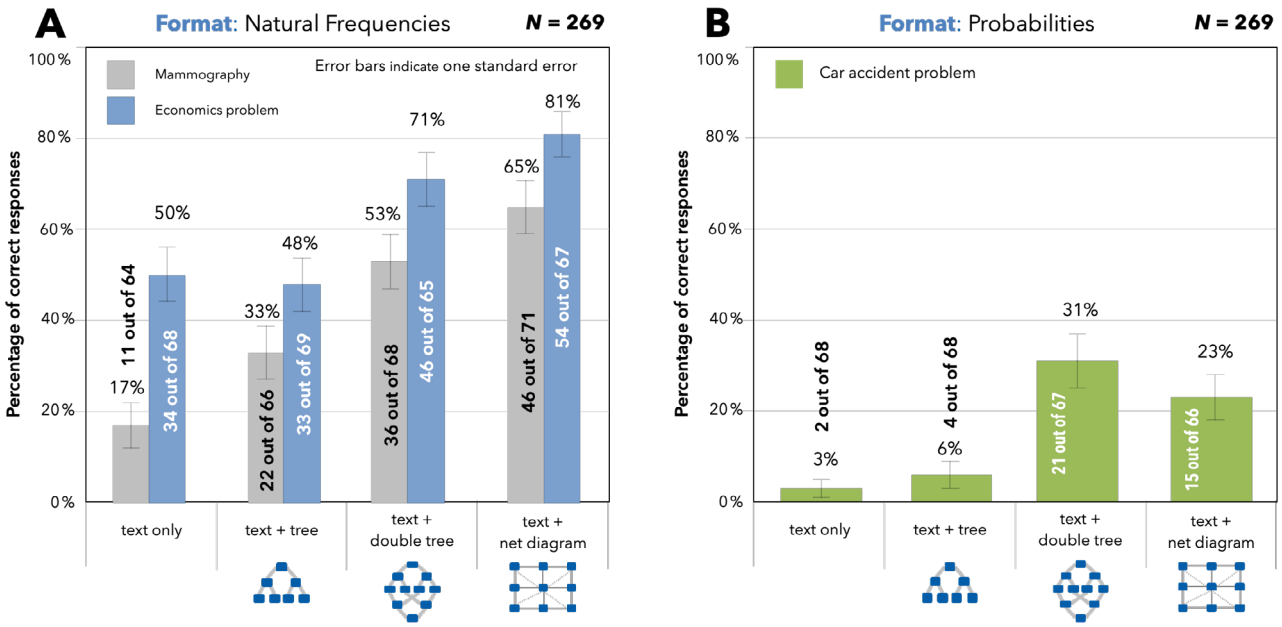


Figure 4: A: Percentages of correct inferences using the different visualizations in the natural-frequency format. B: Percentages of correct inferences using the different visualizations in the probability format.

Study 2–Performance and typical errors with different visualizations

The second study (Binder, Krauss & Wiesner, 2020) focused on the effects of information format (probabilities vs. frequencies) and various graphical models (text only vs. 2×2 table vs. double tree vs. net diagram) on the ability of participants to solve a conditional probability task and a joint probability task. The study also examined the effect of the three visualizations (again depending on information format) on specific student errors.

Information format	Type of question	Conditional probability	Joint probability
	Probabilities	• Bayesian text • 2×2 table • double tree • net diagram	• Bayesian text • 2×2 table • double tree • net diagram
	Natural frequencies	• Bayesian text • 2×2 table • double tree • net diagram	• Bayesian text • 2×2 table • double tree • net diagram

Table 2: Design of the 16 versions implemented in the study. Each student worked on two questions: a conditional probability question and a joint probability question, in the same format (probability or natural frequency).

Method

In a paper-and-pencil study, 249 university students answered questions about conditional probabilities and joint probabilities in typical Bayesian reasoning tasks. The visualizations (2×2 tables, double trees, or net diagrams) were already completely filled in, either in the probability format or in the natural-frequency format. The 16 different versions implemented in the study are described in Table 2.

Results

As can be seen in Figure 5A—results from the conditional probability problem—students performed better overall with natural frequencies (58% correct inferences across visualizations) than with probabilities (23% correct inferences across visualizations). Those participants who used the natural-frequency format performed similarly well with the net (61% correct responses) as with a double tree (60% correct responses)—when asked to work with a visualization that was already fully completed. However, they performed best with the help of a completely filled 2×2 table (78% correct responses). Text-only versions yielded the lowest performance.

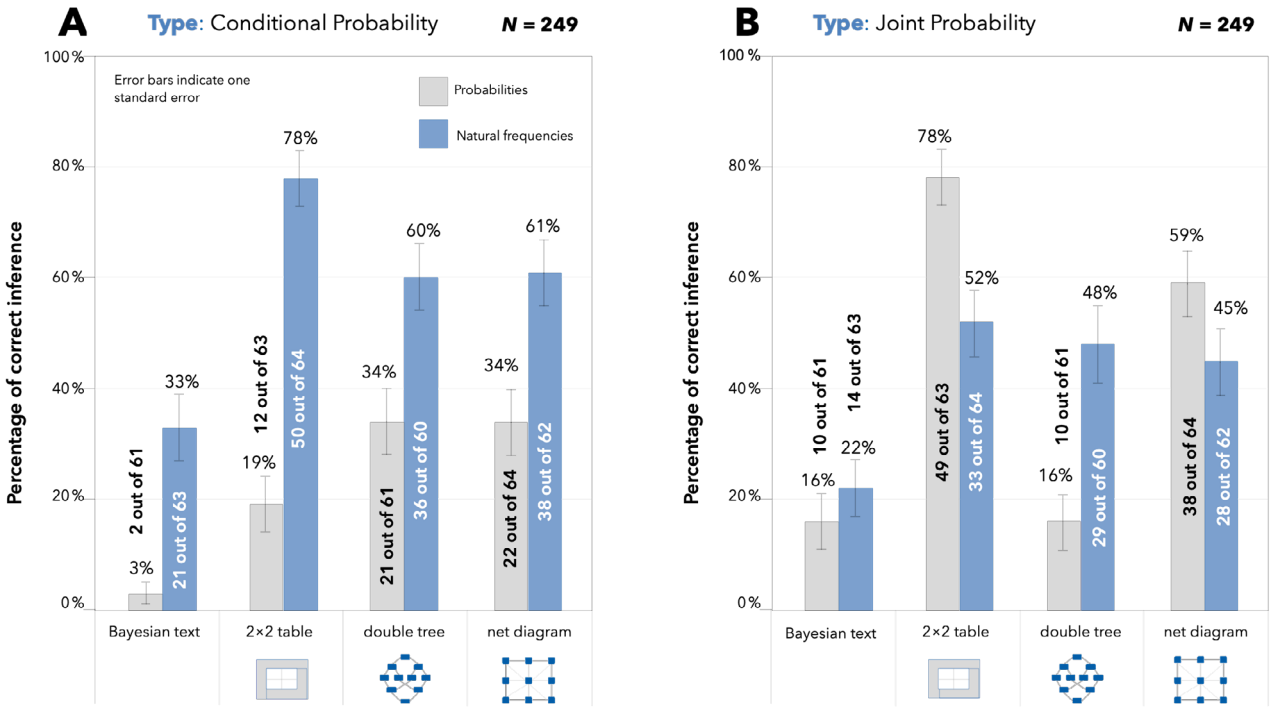


Figure 5: A: Percentages of correct inferences when asking for a conditional probability. B: Percentages of correct inferences when asking for a joint probability.

Figure 5B shows the study results for the joint probability question (e.g., the probability of testing positive *and* having the disease). Interestingly, even though the natural-frequency format has shown consistent advantages in numerous studies about *conditional* probabilities, this study showed no obvious systematic advantage to using natural frequencies for questions about *joint* probabilities. All three visualizations in the natural-frequency format yielded similar performance rates. However, using the probability format, the completely filled net diagram (59% correct inferences) and the completely filled 2×2 table (78% correct inferences) supported participants better in their decision making processes.

An analysis of specific student errors showed interesting shifts between the different information formats, and between the different visualizations. The joint occurrence error (confusion of $P(A|B)$ with $P(A \cap B)$, e.g., the confusion of the *probability of being infected, if the test is positive* with the *probability of being infected and testing positive*) predominantly occurred in the probability 2×2 table and the probability net, whereas the Fisherian error (confusion of $P(A|B)$ with $P(B|A)$) predominantly occurred in the Bayesian text versions, which might be due to an "observability heuristic," possibly meaning that participants tend to use a number they see directly in the visualization as a solution instead of calculating the correct number (Tversky & Kahneman, 1973). A detailed error analysis can be found in Binder et al. (2020). At the end of the chapter, we discuss how these errors can possibly be avoided by not only focusing on different visualizations as *graphic models*, but also illuminating the *data modeling* aspect.

Conclusion: Comparison of the results of both studies

The two studies presented indicate that the frequency net might be used effectively as a *graphical model* in the classroom and should be investigated in more detail in future studies. Despite the fact that the net diagram showed more information that was irrelevant to the question, participants answered conditional probability questions equally well as with the double tree. In the probability format, both studies showed superiority of the double tree and net diagram over the text variant (and also that double tree and net diagram outperform the tree and 2×2 table). The natural-frequency format also showed, in both studies, that participants performed similarly well with double tree and net diagram. However, in Binder et al. (2020) the frequency 2×2 table outperformed these two graphical models in questions for conditional probabilities.

Overall, it can be stated that reasoning with models in Bayesian situations also depends on the graphical model chosen. While certain graphical models tend to emphasize joint probabilities, other graphical models tend to emphasize conditional probabilities. This also explains the completely different results in Binder et al. (2020) when asked about joint probabilities (instead of conditional probabilities).

In the classroom, therefore, mathematics teachers should be aware of which elements a graphical model highlights in each case and which potential student errors in answering certain questions can be provoked with which specific visualizations. In a productive discussion about the confusion of different probabilities, the frequency net could—from a theoretical point of view—offer a special opportunity, because all probabilities (and thus all typical candidates for confusion) are presented here.

Further ideas from the Minerva School: from graphical models to data modeling

The results of the studies show that different graphical models support students differently. Especially the results on errors from the second study suggest that the main problem is the confusion of different probabilities (i.e. $P(A|B)$, $P(B|A)$ and $P(A \cap B)$). These errors could also be reduced in the classroom by integrating *data modeling* more strongly. Similar to Podworny & Frischemeier (2024, **in this volume on page 15**) or Podworny et al. (2021), who used data cards for creating decision trees, data cards (physical or software-based) could help learners to better identify the set and the subset, which are necessary to come to the correct conditional probability. Binder, Krauss & Wiesner (2020) showed that many of the common mistakes in Bayesian reasoning tasks are due to incorrectly identifying the underlying set, whereas most participants in our study were able to identify the correct *subset*. Therefore, a data-driven approach might be helpful to support learners in identifying the correct set. In the following, a data modeling approach via physical data cards is described to support finding the correct set and subset in questions for conditional probabilities.

Data cards (see Figure 6) for exploring conditional probabilities can show all features of the person on one side of the data cards (also compare Podworny & Frischemeier, 2024 **in this volume on page 15**). Learners could work with the data cards by arranging the cards in a kind of 2×2-table.

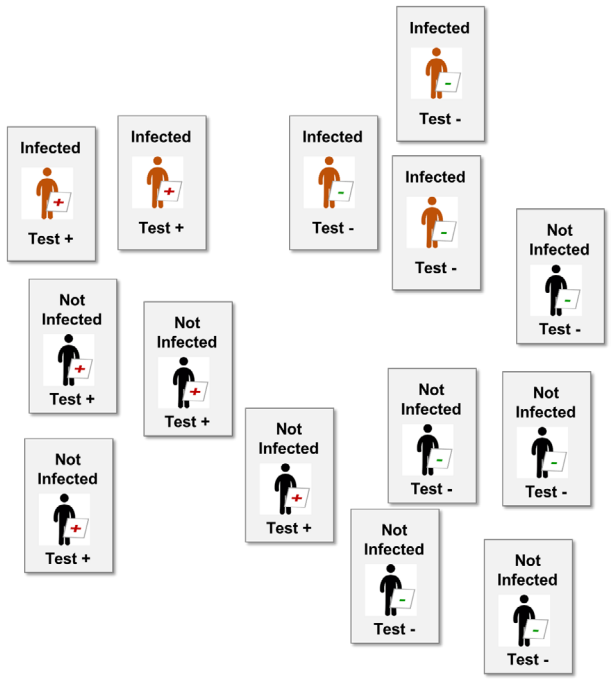


Figure 6: Data cards for conditional probabilities exploration. Both features are visible at the same time.

Addressing several questions regarding different (partial inverted) conditional probabilities should help the learners to distinguish the different linguistic formulations. Lessons on conditional probabilities are always language lessons (Post & Prediger, 2022). This aspect is very important to correctly identify the set and subset. The process of identifying the relevant set and subset can be supported by natural frequencies and therefore also by data cards. If the question is "How many of the persons who are tested positive actually are ill?" do I have to look at all ill persons in the first step or do I have to look at all the people who tested positive? What is the concrete subset here? The natural-frequency format question, but also the process of physically sorting the data cards, might help learners identify the right sets. Future research could explore whether such a sorting process of data cards is helpful to students.

Alternatively, data cards can be structured differently, showing only *one feature per side* of the card, thereby emphasizing the sorting process in a different way. This type of data card is recommended by Fiedler et al. (2000), who suggested a sampling approach to biases in Bayesian reasoning, and compared criterion sampling with predictor sampling (see also Gavanski & Hui, 1992 and Wason, 1966). According to this approach there are basically two ways of sorting the deck of cards (see Figure 7): Structuring the set by infected vs. no infected (see Figure 7A) or structuring the set by test positive vs. test negative (see Figure 7B). Should one choose the data cards on the left side, look at all infected persons and check, which proportion of them is positive tested? Or is it better to choose the data cards on the right side, look at all persons tested positive and check, which proportion of them is infected? This enactive approach, with the manual sorting of the data cards and the conscious choice of one of the two sets of data cards, could be a data-based modeling process that might prevent confusion between conditional probabilities.

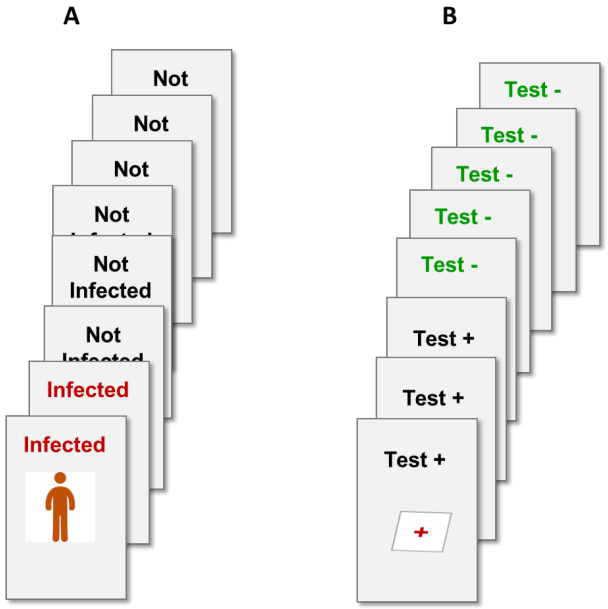


Figure 7: Data cards for conditional probabilities exploration. Only one of the features is directly visible. The other feature is on the back of the data card.

The question that needs to be answered is: How can we enable learners to choose the correct one of the two sets of cards to answer the question "What is the proportion of those who are infected among those who test positive?" These data-based sorting operations might be helpful before finally drawing a graphical model (e.g., a tree diagram or net diagram).

References

Batanero, C., & Sanchez, E. (2013). What is the Nature of High School Students' Conceptions and Misconceptions About Probability? In: Graham A. Jones: *Exploring probability in school: Challenges for teaching and learning*, 260–289, Kluwer Academic Publishers.

Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information: An empirical study on tree diagrams and 2 × 2 tables. *Frontiers in Psychology*, 6(1186).

Binder, K., Krauss, S., & Wiesner, P. (2020). A new visualization for probabilistic situations containing two binary events: the frequency net. *Frontiers in Psychology*, 11(750).

Binder, K., Steib, N., & Krauss, S. (2023). Von Baumdiagrammen über Doppelbäume zu Häufigkeitsnetzen—Kognitive Überlastung oder didaktische Unterstützung? [Moving from tree diagrams to double trees to net diagrams—cognitively overwhelming or educationally supportive?] *Journal für Mathematik-Didaktik*, 44, 471–503.

Böcherer-Linder, K., & Eichler, A. (2019). How to improve performance in Bayesian inference tasks: a comparison of five visualizations. *Frontiers in Psychology*, 10, 267.

Blum, W., Galbraith, P., Henn, H.-W., & Niss, M. (Eds.) (2007). *Modelling and applications in mathematics education. The 14th ICMI study*. New York: Springer.

Eichler, A., & Vogel, M. (2015). Teaching Risk in School. *The Mathematics Enthusiast*, 12(1).

Eichler, A., & Vogel, M. (2013). *Leitidee Daten und Zufall. Sekundarstufe I*. Wiesbaden: Vieweg + Teubner.

Fenton, N. (2011). Improve statistics in court. *Nature*, 479(7371), 36–37.

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General*, 129(3), 399.

Gal, I. (2024). **"What do citizens need to know about real-world statistical models and the teaching of data modeling" in this volume on page 91.**

Gavanski, I., & Hui, C. (1992). Natural sample spaces and uncertain belief. *Journal of Personality and Social Psychology*, 63(5), 766

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, 102(4), 684–704.

Khan, A., Breslav, S., Glueck, M., & Hornbæk, K. (2015). Benefits of visualization in the mammography problem. *International journal of human-computer studies*, 83, 94–113.

Martignon, L., Frischemeier, D., McDowell, M., & Till, C. (2023). Dynamic, Interactive Trees and Icon Arrays for Visualizing Risks in Civic Statistics. In *Statistics for Empowerment and Social Engagement: Teaching Civic Statistics to Develop Informed Citizens* (pp. 477–501). Cham: Springer International Publishing.

McDowell, M., & Jacobs, P. (2017). Meta-Analysis of the Effect of Natural Frequencies on Bayesian Reasoning. *Psychological bulletin*, 143, 1273–1312.

Pfannkuch, M., & Budgett, S. (2017). Reasoning from an Eikosogram: An exploratory study. *International Journal of Research in Undergraduate Mathematics Education*, 3(2), 283–310.

Podworny, S., & Frischemeier, S. (2024). **"Young learners' perspectives on the concept of data as a model: what are data and what are they used for?" in this volume on page 15.**

Podworny, S., Fleischer, Y., Hüsing, S., Biehler, R., Frischemeier, D., Höper, L., & Schulte, C. (2021, November). Using data cards for teaching data based decision trees in middle school. In *Proceedings of the 21st Koli Calling International Conference on Computing Education Research* (pp. 1–3).

Post, M., & Prediger, S. (2022). Teaching practices for unfolding information and connecting multiple representations: the case of conditional probability information. *Mathematics Education Research Journal*, 1–33.

Stine, G. J. (1996). *Acquired immune deficiency syndrome: Biological, medical, social, and legal issues*. Englewood Cliff, NJ: Prentice Hall.

Tversky, A. & Kahneman, D. (1973) Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 1973, S. 207–232.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Hamondsworth, England: Penguin.

Wassner, C., Martignon, L., & Biehler, R. (2004). *Bayesianisches Denken in der Schule* [Bayesian reasoning in schools]. Unterrichtswissenschaft, 32(1), 58–96.

Wegwarth, O., & Gigerenzer, G. (2013). Overdiagnosis and overtreatment: evaluation of what physicians tell their patients about screening harms. *JAMA internal medicine*, 173(22), 2086–2088.

Supporting students' modeling and data practices by engaging with digital tools

Tom Bielik

Beit Berl College, Israel
tom.bielik@beitberl.ac.il

Modeling is a key scientific and engineering practice. Scientists develop and use models to communicate and critique their ideas. Students are expected to engage in the modeling practices of developing, using, and revising their models in science classroom, together with data practices of collecting, analyzing, and communicating data. Supporting students' modeling and data practices can advance their systems and computational thinking skills when solving problems and making sense of complex phenomena. However, students lack opportunities to meaningfully engage in modeling and data practices in science classrooms. In this chapter, I focus on the affordances, opportunities, and challenges students face when engaging with SageModeler, a computational modeling tool, as reflected from four studies in recent years and in relation to the system modeling practices theoretical framework presented in the literature review below (Bielik et al., 2019). Findings suggest that engaging with the computational modeling tool can be challenging for students, but these activities can develop their modeling and data practices, complex systems understanding, and metamodeling knowledge. I discuss these findings in connection to other chapters in this book and conclude with several recommendations for researchers, educators, and curriculum designers.

Introduction

Preparing students to be scientifically literate 21st century citizens is a major educational challenge, especially when facing today's increasing need for critical and creative thinking in the technology-rich job market (OECD, 2018). A Framework for K–12 Science Education (NRC, 2012) calls for shifting the focus from learning about scientific principles to making sense of phenomena and problem-solving by engaging students with scientific and engineering practices and crosscutting concepts, such as modeling and systems models. This call is enhanced in face of global challenges in recent years such as climate change and COVID-19 pandemic.

Providing students with meaningful opportunities to engage with authentic scientific practices can improve their learning achievements and interest in science (NRC, 2012). Modeling is a key science and engineering practice, serving as an epistemic tool employed by scientists to represent their ideas, to engage in scientific inquiry, and to communicate their ideas (Harrison & Treagust, 2000; Louce & Zacharia, 2012). It can also support the development of students' metamodeling knowledge, which refers to knowledge about the nature, purpose and process of scientific modeling (Göhner et al., 2022; Schwarz et al., 2009). Models are constructed and tested using data, either collected by the students themselves or provided to them as secondary sources. Using data requires students to develop data practices, such as organizing, sorting, and analyzing data (Pfannkuch et al., 2018). Therefore, students are expected to construct, use, test, and revise models while collecting, analyzing, and representing data in classroom. However, most students do not have meaningful opportunities to develop their modeling practices (Schwarz et al., 2009).

This chapter includes a summary of results from four empirical studies carried out by our research group and discusses how these studies relate to the system modeling practices theoretical framework presented by Bielik, Stephens, Damelin, and Krajcik (2019). These studies present the implementation of middle- and high-school curricular units using SageModeler, a computational modeling tool. Bielik, Opitz, and Novak (2018) focus on the implementation of a 7th grade unit about water quality in a local watershed. Bielik, Damelin, and Krajcik (2019) focus on the enactment of a 7th grade unit about ocean acidification, which included real-world big data analysis. Bielik, Fonio, Feinerman, Golan Duncan, and Levy (2020) focus on the implementation of a 9th grade unit about ant behavior, in which several modeling tools were incorporated. Finally, Bielik, Stephens, McIntyre, Damelin, and Krajcik (2021) provides results from enactment of a 10th grade chemistry unit about the ideal gas law. In the discussion, I reflect on the affordances and challenges students face when developing and using digital modeling tools such as SageModeler and present recommendations for advancing students' engagement with computational modeling tools.

Literature review

Models and modeling in science education

Scientific modeling (hereafter referred to as "modeling") is a key scientific and engineering practice emphasized in the latest science education science standards (NRC, 2012), in which students are expected to develop, use, and revise their models in the science classroom (Harrison & Treagust, 2000; Nersessian, 2002; Passmore et al., 2014). Scientific models are broadly defined as epistemic tools that are used to explain and predict phenomena, composed of system components and the relationships between them. Engaging students in modeling should build their cognitive and epistemic scientific knowledge and understanding (Schwarz et al., 2009).

The goal of modeling is to test ideas by representing systems of connected processes and evaluating them with real-world data (Passmore et al., 2014; Windschitl et al., 2008). Students who are provided with meaningful opportunities to engage in modeling develop their epistemic understanding about scientific models (i.e., metamodeling knowledge), which includes understanding that models serve as a tool for thinking about systems rather than object description, that models are never complete, and that they represent the current consensus understanding based on known empirical evidence (Göhner et al., 2022). Students can best learn about models and modeling when provided with activities that build on their prior knowledge. However, students require substantial support—including through social negotiation, see Harrison & Treagust, 2000—and repeated experiences to fully develop their modeling practices.

In traditional science classrooms, students usually do not have meaningful opportunities to engage in modeling, and both teachers and students often lack understanding of modeling practices (Schwarz, 2009; Windschitl et al. 2008). Often, teachers fail to stress the limitations of models and assume students understand that models are always tested and revised. Most students view scientific models as realistic algorithmic representations that could be used to memorize the correct answer, rather than as epistemic explanatory and inquiry tools for sharing and critiquing ideas (Harrison & Treagust, 2000).

Data modeling and digital tools

Data plays a key role in modeling, as discussed by Susanne Podworny and Daniel Frischemeier's chapter (**in this volume on page 15**). Data modeling is an essential part of the modeling process, where students are expected to engage with self-produced or secondary data to develop and test their models (Berland et al., 2016; Chinn & Brewer, 2001; Weintrop et al., 2016). This requires them to develop data practices, such as sorting, organizing, and analyzing data to be used in the model, together with statistical thinking (Pfannkuch et al., 2018).

Digital tools can support students' modeling and data practices by providing them with opportunities to build computational models, run simulations of their data, and use data to test, evaluate and revise their models (NRC, 2012; Weintrop et al., 2016). As discussed in the chapter of Christian Büscher (**in this volume on page 49**), digital tools can also support students' statistical literacy. Computational modeling tools can be particularly effective in supporting students, as they give students an opportunity to explore complex dynamic relationships between components in the model and to visualize abstract concepts (Crawford and Cullin, 2004; Louca and Zacharia, 2012; Shin et al., 2022). Computational models are also useful when analyzing, abstracting, and recognizing patterns in big data (Grover & Pea, 2018; Weintrop et al., 2016; Wing, 2014). Integrating digital tools in learning environments requires addressing systemic issues such as usability, scalability, and sustainability of the tool to make their use widespread in science classrooms (Fishman et al., 2004), and more research is needed to explore students' learning with and about models and to investigate the possible effect of model-based teaching on students' conceptual understanding and development of their metamodeling knowledge.

Bielik, Stephens, Damelin, and Krajcik (2019) presented a theoretical framework that included four aspects of systems modeling practices. These aspects focus on the core elements that are essential for modeling when using computational models to make sense of phenomena. It is based on the core modeling practices of constructing, testing, revising, and using models (Schwarz et al., 2009). The four aspects are:

1. defining the boundaries of the system by including components in the model that are relevant to the phenomenon under investigation;
2. determining appropriate relationships between components in the model;
3. using evidence and reasoning to construct, use, evaluate, and revise the model; and
4. interpreting the behavior of the model to determine its usefulness in explaining and making predictions about phenomena.

In the studies presented below, I present the main findings from our studies considering the system modeling practices framework presented above. I discuss the affordances and challenges students face when engaging with digital computational modeling tools in school-science curricular units.

Methodology

Context: the SageModeler computational modeling tool

The computational modeling tool, SageModeler, was designed to support students in constructing and using models to explain phenomena and design solutions using a range of visual representations such as images, text, labels, tables, and graphs (Damelin et al., 2017). Using the tool, students create models portraying relationships between variables and run simulations to explain and predict phenomena. The variables, represented by student-selected images and labels, are connected by arrows. The causal relationships between the variables are defined using semi-quantitative descriptions. The modeler chooses the direction of the relationship (e.g., "increase", or "decrease") and the magnitude of the relationship (e.g., linear relationship as "about the same" or "a lot" or logarithmic relationship as "more and more"). The choice of relationship magnitude is accompanied by a graphical illustration, which provides students with an opportunity to visualize the mathematical relationship. Figure 1 presents an example of a SageModeler model about photosynthesis which includes the model variables and relationships, relationship definition box, and output graph and table.

SageModeler models can be designed as "static equilibrium simulations", which consist of a set of variables and relationships without representing change over time. This allows the modeler, for example, to develop experimental models that show the effect of different independent variables on the dependent variable. Another option is to design a "dynamic time-based simulation" model, which allows the modeler to simulate accumulation or transfer of components over time. In these "stocks and flows" type of models, the modeler can simulate complex phenomena such as energy transfer in physical systems or flows of matter in ecosystems (Eidin et al., 2023). Examples of both type of simulations can be found in SageModeler settings.

After setting up the variables and defining relationships, students can run quantitative simulations in which they manipulate the independent variables to receive output as graphs or tables. Students can test and revise their models while working with empirical data from different sources, such as classroom experiments or secondary authentic research data. The tool was developed to support students' modeling practices, based on learning progressions described in *A Framework for K–12 Science Education* and empirical research (NRC, 2012), and designed to be easy to use, intuitive, interactive, and visually engaging. SageModeler is integrated into the Common Online Data Analysis Platform (CODAP), a graphing and data analysis platform that takes the outputs generated by the model and any other data source to combine them into a single analytic environment (Finzer & Damelin, 2016). Importing authentic experimental data into CODAP allows students to compare, evaluate, and revise their model to better fit real-world evidence. Students can create multiple graphs from these data sources and use them to test the validity of their models.

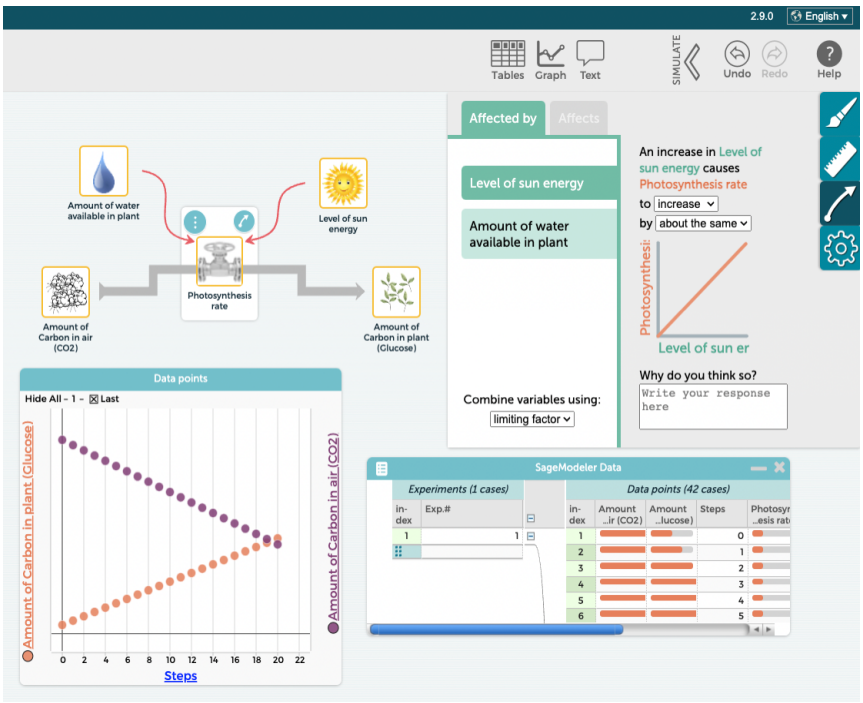


Figure 1. An example of a SageModeler dynamic time-based simulation model showing variables that affect the photosynthesis rate in plants. The model itself is located at the top left, the relationship definition box is at the top right, and the output graph and table are at the bottom.

Tools and methods

In the four studies presented in this paper, both qualitative and quantitative methods were used in a design-based research approach that examined the implementation of curricular units that included computational modeling tools in secondary science classrooms. Collected data included students' produced models and artifacts, video, audio, and screencast recordings of the implemented lessons, pre- and post-questionnaires evaluating students' content learning and metamodeling knowledge, and interviews with teachers and students.

Results

This section includes brief descriptions of the most salient results from four separate studies, condensed to focus on this chapter's topic. For additional details, please refer to the source papers for each study.

Students' engagement with the modeling practices

In Bielik, Opitz, and Novak (2018), we investigated how modeling practices, i.e., constructing, using, evaluating, and revising models, were integrated into a middle school curricular unit about water quality that included using SageModeler. This was a design-based study that included a qualitative analysis of classroom observations, video recordings, teacher reflection notes, and pre- and post-enactment modeling surveys. The study focused on three student groups as case studies to track the development of students' modeling practices and metamodeling knowledge across the unit. Students constructed, used, evaluated, and revised their models based on data they collected and analyzed from a local watershed system of connected water ponds in their school and connected it with scientific concepts they learned during the unit related to watersheds and water quality. Results indicated that most students succeeded in constructing appropriate complex models of the water quality in the local watershed using the modeling tool by adding and specifying variables and relationships to their models. We also saw development of basic metamodeling concepts regarding the representational properties of models. Classroom observations and recordings showed that most students engaged, to some extent, with all **four aspects** of system modeling practices: when choosing the relevant variables to include in their models (aspect 1), when determining the direction and magnitude of relationships between the variables (aspect 2), when integrating evidence and data from their performed experiments to construct, evaluate and revise their models (aspect 3), and when running their models to simulate and explain the investigated water quality phenomenon in whole class presentations (aspect 4).

In all case studies examined in this study, students' models progressed in their complexity by integrating additional variables at each modeling cycle to improve the model's explanatory and predictive power and to fully address the driving question of the unit. Students' models also progressed in their quality, though to a lesser extent compared to the progress in its complexity. This indicated that students focused more on adding variables and relationships to their models in each revision cycle, rather than in re-examining the correctness of the existing variables and relationships based on the data they collected. These findings pointed towards progress in the performance of students' modeling practices. Less progress was found in students' metamodeling knowledge. Results also suggested that students faced several challenges when using the computational modeling tool and developing their modeling practices, mostly when evaluating and revising their models. Some student groups were not able to identify and correct mistakes in their models, as they mostly focused on adding new variables when revising their models. These mistakes included undefined relationships between variables in the model, inaccurate relationships between variables in terms of directionality or magnitude, missing variables, and more. Our results suggested that using a modeling tool can support students' modeling practices and that repeated opportunities for students to evaluate and revise their computational models can improve their complexity and quality.

Students' engagement with data modeling

In Bielik, Damelin, and Krajcik (2019), we investigated the integration of the modeling tool in a 7th grade middle school unit focusing on ocean acidification. The unit included research using a big data set on the environmental conditions at the Aloha Station in Hawaii. This was a design-based study that included analysis of models students produced, student and teacher interviews, pre- and post-enactment surveys, and students' responses to "question-at-the-door" surveys at the end of lessons. Students developed static equilibrium simulation models based on data they collected in a set of experiments they performed, information they collected online, and a big data set they were provided. Students had the opportunity to explore the big data set and to create graphs that show trends found in the data. We investigated the advantages and challenges experienced by students and teachers while engaging in the unit and while using the modeling tool. Results indicated that integrating the modeling tool in the unit facilitated students' interest and engagement, developed their sense of environmental responsibility, and focused their attention toward human involvement and impact on the environment. Students successfully engaged with the big data set as secondary data. They used the data to produce graphical representations and to test their models. Students perceived the modeling tool and the curricular unit to be relevant to their lives and important in promoting their content learning and modeling practices. In this enactment, as in the study described above, most students had successful engagement with all aspects of system modeling practices.

Students and teachers also reported facing several challenges, mostly related to the complexity of using the modeling tool, working with big data, and producing the graphs and charts. Students were provided with opportunities to engage with all the modeling practices while developing their models in an iterative process. They

continuously developed and revised their models as they broadened their understanding about ocean acidification during the investigations carried out in the lessons. Our findings suggested that the process of iteratively engaging with computational models can contribute to students' affective, cognitive, and behavioral engagement, as reported by the teachers and students in this study and as was found in the artifacts of the unit. However, the constraints and limitations mentioned by the teachers and students should be considered when designing curricular units that include computational modeling tools.

Students' engagement with different modeling approaches

In Bielik, Fonio, Feinerman, Golan Duncan, and Levy (2020), we described the design principles used to develop and implement a 9th grade curricular unit about ants' collective behavior that integrates three modeling approaches:

- Conceptual "drawn" models, i.e., diagrams on paper.
- Agent-based computational models (ABM) that focus on the aggregate macro-level behavior emerging from the micro-level behavior of the agents in the model (Thompson & Reimann, 2010). Here, students used a NetLogo simulation of the ants.
- System dynamics computational models (SD), focusing on the complex non-linear, and feed-back-loop characteristics of phenomena (Russ et al., 2008). In this study, this is the student work in SageModeler.

This was a qualitative study that included analysis of models students produced, lesson recordings and observations, pre- and post-enactment questionnaires, and a teacher interview. Students were provided with a partially developed dynamic time-based SageModeler model that included some of the factors affecting the ants' rate of food transfer from the environment to the nest. They were directed to complete the model based on their investigation using the agent-based computational simulation. The three modeling approaches (conceptual models, ABM, and SD) provided students with opportunities to share their ideas about ants' collective behavior and to investigate the factors that influence the efficiency of ant food foraging. Students' knowledge about ants' behavior developed following learning the unit. Some gains were also found in students' metamodeling knowledge about models as tools for investigating phenomena. However, no significant changes were found in students' perception of models as tools for explaining and predicting phenomena.

The collaborative nature of the activities in the ant-behavior unit was an important factor that pushed students to fully engage with the modeling tools, since in all lessons students worked in small groups to develop, test, and use their models and to share their models with their peers to receive feedback. However, students required technical and conceptual support when using the computational modeling tools, since this was a new practice for them, and they did not have many opportunities to use such digital tools in school prior to the intervention. In this enactment, students had somewhat limited engagement with the first two aspects of system modeling practices (constructing and using models) since they were provided with partially built computational models and defined only some of the relationships in the SageModeler model. However, they showed high engagement with the other two aspects, as they used evidence obtained from the agent-based modeling tool to test the relationships in the SageModeler model and used the simulation results to explain the phenomenon of ant communication and collaboration when foraging for food.

Students' engagement with the four aspects of systems modelling practices

Finally, in Bielik, Stephens, McIntyre, Damelin, and Krajcik (2021), we studied the enactment of a system modeling chemistry unit focusing on the emergent properties of gases. We examined evidence of 10th grade students' engagement with the four aspects of systems modeling practices in depth. This was a design-based study that included analysis of students' model reflection questions integrated in the online activities, students' produced models, and student interviews. Students developed dynamic time-based simulation models showing the factors that caused a real-life phenomenon of a large tank implosion, based on a set of experiments and investigations they performed. We explored the choices students made when constructing their models, whether they described evidence and reasoning for those choices, and whether they examined the behavior of their models in connection with model usefulness in explaining and making predictions about the phenomena of interest.

In this study, we found that students' engagement with system models mostly increased near the end of the unit when this information helped them evaluate and revise their models. In addition, students required different kinds of scaffolding support in how to use real world data to help them improve their models. Students progressed in their ability to choose appropriate variables, determine relevant relationships, and clarify causal mechanisms to make the relationships in their models more elaborated. Less progress was observed in respect to using data and evidence to support model design and explicitly linking the overall behavior of the model to the driving question about the phenomenon under investigation. Students experienced several challenges with causal reasoning,

including providing evidence and reasoning for their chosen variables and relationships in the models, and explaining how their models addressed the driving question of the unit.

Discussion

The results described above indicate that using digital modeling tools and big data, alongside supportive curricular materials that focus on the development of students' modeling and data practices, can support students' engagement, conceptual learning, and meta-modeling knowledge in school science lessons. In the lessons, students developed and revised their models and used them to run simulations and make predictions of the investigated phenomena. These studies provide insights into how students engage in modeling and data practices in science classrooms using a computational modeling tool. These studies also highlight the main characteristics and learning achievements of students when reasoning with models and using big data in digital tools, and suggest practical approaches in which modeling and data practices can be used for scientific inquiry and reasoning.

In the studies presented in this chapter, students required technical and cognitive supports when using computational modeling tools, as the modeling and data practices were unfamiliar to them, and they did not have many opportunities to use such digital tools in school prior to the interventions. Sufficient time to meaningfully engage with the models and data was one of the main limiting factors during the implemented lessons, resulting in few opportunities to engage students in broader discussions about the affordances and constraints of the modeling process and the nature of models. These findings also align with the challenges students face when engaging with mathematical modeling and hands-on experiments, as discussed by Ramona Hagenkötter and her colleagues (**in this volume on page 41**). It is believed that integrating several different modeling approaches holds a strong promise to promote science students' learning, provided they are given appropriate scaffolds and sufficient time to engage with each of the modeling approaches and to discuss the affordances and limitations of each. Selecting the appropriate pedagogies for supporting students in developing data-based models is a major challenge for 21st-century citizens literacy, as discussed in Iddo Gal's chapter (**in this volume on page 91**). On top of that, as discussed by Robert Gould (**in this volume on page 81**), the modeling approach and tool must match the purpose of the model. SageModeler, like any other computational modeling tool, should be used for the appropriate purpose and learning goal. A curriculum that integrates modeling tools should be carefully designed to account for students' cognitive level and content knowledge, and facilitate their modeling and data practices. Classroom implementation results suggest that identifying appropriate curricular activities and teacher supports are key for learning and the development of students' modeling and data practices.

Implications and contribution

The findings described in this chapter align with the 2022 Minerva school theme, mostly in the aspects of learning environments that foster reasoning with data models and designing modeling-centered pedagogies. It holds the potential to support researchers, educators, and curriculum designers interested in integrating digital tools to support students' modeling and data practices. My experiences at the 2022 Minerva school provided me with more insightful ideas and research directions to integrate modeling and data practices in school science activities that can further support students' engagement and learning.

As indicated in the studies presented in this chapter and in the literature, supporting the development of students' modeling and data practices using digital tools requires continued feedback and interactions between the students and the teacher with scaffolding supports from the curricular materials. Modeling tools such as SageModeler can support the students in the modeling process, while making it more interactive, dynamic, and engaging.

Based on the results from the studies described above and in alignment with the four aspects of the system modeling practices framework, the following recommendations are suggested for science educators and curriculum designers interested in developing and supporting students' modeling and data practices using computational modeling tools:

1. Focus on using data and evidence to support critical evaluation of model components and the relationships between them. Running simulations to evaluate the outcome of a model in comparison with real-life data is an important feature when using digital modeling tools. As seen in Bielik, Opitz, and Novak (2018), some students face difficulties when addressing problems in their models during model revision activities. This is a crucial checkpoint in the modeling process, and teachers should direct students to carefully test their models throughout the model development process. In addition, as found in Bielik, Stephens, McIntyre, Damelin, and Krajcik (2021), students require different scaffolds from the teachers and curricular materials to support them in using data to evaluate their models.

2. Evaluate models in whole-class and small-group discussions. All studies presented in this chapter were designed as collaborative inquiry units that include many opportunities for students to discuss and present their ideas using the computational modeling tools. This was mostly emphasized in Bielik, Fonio, Feinerman, Golan Duncan, and Levy (2020), where students had repeated opportunities to share and discuss their models throughout the lessons. Getting students to talk through their models can be helpful in identifying inconsistencies in their models and inappropriate model behavior. Student-centered discussions are powerful tools for sharing ideas related to the phenomena being modeled and for engaging students in activities that support growth in metamodeling knowledge. These discussions can include peer review, gallery walks, presentations, and collaborative evaluation of student models.
3. Frequently revisit the overarching phenomenon and the goal that the model is intended to achieve. Students can easily lose the big picture of what they are modeling and why they analyze the data, especially when using cognitively demanding digital tools, as observed in Bielik, Damelin, and Krajcik (2019). Teachers should frequently emphasize and revisit the goal of model development and respond to student questions and comments related to it. It is also suggested that the metamodeling knowledge about the purpose of modeling should be consistently visible for the students while they develop and use their model.

References

Berland, L. K., Schwarz, C. V., Krist, C., Kenyon, L., Lo, A. S., & Reiser, B. J. (2016). Epistemologies in practice: Making scientific practices meaningful for students. *Journal of Research in Science Teaching*, 53(7), 1082–1112.

Bielik, T., Damelin, D., & Krajcik, J. (2019). Shifting the Balance: Engaging Students in Using a Modeling Tool to Learn about Ocean Acidification. *Eurasia Journal of Mathematics, Science and Technology Education*, 15(1).

Bielik T., Fonio E., Feinerman, O., Golan R. T., & Levy S. T. (2021). Working Together: Integrating Computational Modeling Approaches to Investigate Complex Systems. *Journal of Science Education and Technology*, 30 (40–57).

Bielik T., Opitz S., & Novak M. A. (2018). Supporting Students in Building and Using Models: Development on the Quality and Complexity Dimensions. *Education Sciences*, 8(3), 149.

Bielik T., Stephens L., Damelin D., & Krajcik J. Designing Technology Rich Environments to Support Student Modeling Practice (2019). In Upmeir Zu B., Kruger D., & Van Driel J. (Eds.), *Towards a Competence-based View on Models and Modeling in Science Education*. Springer International Publishing (275–290).

Bielik T., Stephens L., McIntyre C., Damelin D., & Krajcik J. (2021) Supporting Student System Modeling Practice Through Curriculum and Technology Design. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-021-09943-y

Büscher, C. **"Design principles for developing statistical literacy by integrating data, models, and context in a digital learning environment" in this volume on page 49.**

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction*, 19(3), 323–393.

Crawford, B.A.; Cullin, M.J. (2004). Supporting prospective teachers' conceptions of modelling in science. *International Journal of Science Education*, 26, 1379–1401.

Damelin D., Krajcik J., Mcintyre C., & Bielik T. (2017). Students Making Systems Models: An Accessible Approach. *Science Scope*, Vol. 40.5, 78–82.

Eidin E., Bielik T., Touitou I., Bowers J., McIntyre C., Damelin D., & Krajcik J. (2023) Thinking in Terms of Change over Time: Opportunities and Challenges. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-023-10047-y

Finzer, W., & Damelin, D. (2016). Design perspective on the Common Online Data Analysis Platform. In C. E. Konold (Chair), *Student thinking, learning, and inquiry with the Common Online Data Analysis Platform*. Symposium conducted at the meeting of the American Educational Research Association, Washington, D.C.

Fishman, B., Marx, R. W., Blumenfeld, P., Krajcik, J., & Soloway, E. (2004). Creating a framework for research on systemic technology innovations. *The Journal of the Learning Sciences*, 13(1), 43–76.

Gal, I. **"What do citizens need to know about real-world statistical models and the teaching of data modeling" in this volume on page 91.**

Göhner, M. F., Bielik, T., & Krell, M. (2022). Investigating the dimensions of modeling competence among preservice science teachers: Meta-modeling knowledge, modeling practice, and modeling product. *Journal of Research in Science Teaching*, 59(8), 1354–1387.

Gould, R. **"Traditional statistical models in a sea of data: teaching introductory data science" in this volume on page 81.**

Grover, S., & Pea, R. (2018). Computational thinking: A competency whose time has come. *Computer science education: Perspectives on teaching and learning in school*, 19(1), 19–38.

Hagenkötter, R., Nachtigall, V., Rolka, K. & Rummel, N. **"Mathematical hands-on experimentation as a possibility to engage students in authentic modeling with real data" in this volume on page 41.**

Harrison, A. G., & Treagust, D. F. (2000). A typology of school science models. *International Journal of Science Education*, 22(9), 1011–1026.

Louca, L. T., & Zacharia, Z. C. (2012). Modeling-based learning in science education: Cognitive, metacognitive, social, material and epistemological contributions. *Educational Review*, 64(4), 471–492.

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Nersessian, N. J. (2002). The cognitive basis of model-based reasoning in science. In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The cognitive basis of science* (pp. 133–153). Cambridge: Cambridge University Press.

OECD. (2018). *The future of education and skills: Education 2030*. OECD Education 2030.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., ... & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, 46(6), 632–654.

Shin, N., Bowers, J., Roderick, S., McIntyre, C., Stephens, A. L., Eidin, E., ... & Damelin, D. (2022). A framework for supporting systems thinking and computational thinking through constructing models. *Instructional Science*, 50(6), 933–960.

Passmore, C., Gouvea, J. S., & Giere, R. (2014). Models in science and in learning science: Focusing scientific practice on sense-making. In M. R. Matthews (Ed.), *International handbook of research in history, philosophy and science teaching* (pp. 1171–1202). Netherlands: Springer.

Podworny, S. and Frischemeier, D. **"Young learners' perspectives on the concept of data as a model: what are data and what are they used for?" in this volume on page 15.**

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

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: a framework for discourse analysis developed from philosophy of science. *Science Education*, 92(3), 499–525.

Thompson, K., & Reimann, P. (2010). Patterns of use of an agent-based model and a system dynamics model: the application of patterns of use and the impacts on learning outcomes. *Computers & Education*, 54(2), 392–403.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology*, 25(1), 127–147.

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, 92(5), 941–967.

Wing, J (2014). Computational Thinking Benefits Society. *Social Issues in Computing*. Available at: http://socialissues.cs.toronto.edu/2014/01/computational- thinking/

Mathematical hands-on experimentation as a possibility to engage students in authentic modeling with real data

Ramona Hagenkötter, Valentina Nachtigall, Katrin Rolka, and Nikol Rummel

Ruhr-University Bochum

ramona.hagenkoetter@rub.de, valentina.nachtigall@rub.de, katrin.rolka@rub.de, nikol.rummel@rub.de

In the big data era, people need to become active data explorers who understand how data can be used to describe and model the world. Therefore, students should acquire key data literacy competencies already in school. Against this background, the present chapter introduces mathematical modeling with hands-on experimentation as one potentially promising approach to engage students in authentic modeling with real data, and thereby, promote their data literacy as well as their conceptions about mathematics. As findings of our interview study suggest, students often do not associate mathematics with processes of active data exploration but instead with mere schematic-algorithmic application of predefined steps. Moreover, insights into students' mathematical modeling with real data from mathematical hands-on experimentation suggest that students need to learn to mathematically model real-world phenomena and, especially, to deal with real data. Building on these findings, the present chapter calls for the implementation of mathematical modeling with hands-on experimentation in mathematics education to provide students with more opportunities to deal with real data and to become active data explorers.

Introduction

It is important in the big data era that people are not simply passive recipients of data-based reports. Rather, they need to become active data explorers who can "identify, collect, evaluate, analyze, interpret, present, and protect data" (Oceans of Data Institute, 2015, p. 4). Therefore, every citizen needs to further develop data-related skills which are referred to as *data literacy*. According to Risdale et al. (2015, p. 8), data literacy is "the ability to collect, manage, evaluate, and apply data, in a critical manner" and is essentially required in the global knowledge-based economy. In addition, Wolff et al. (2015, p. 23) analyzed different perspectives on data literacy and surveyed existing approaches to teaching data literacy in practice. Based on their findings, they describe data literacy as "the ability to ask and answer real-world questions from larger and small data sets through an inquiry process, with consideration of ethical use of data." Moreover, Wolff et al. (2015) argue that the foundation for a data literate society begins by acquiring key data literacy competencies in school. Therefore, mathematics classes can be considered as one opportunity to foster such data literacy competencies. However, as indicated by our own research (Hagenkötter et al., 2022) as well as previous studies on students' conceptions about mathematics in general (e.g., Schoenfeld, 1992), students in mathematics classes are often only passive consumers of others' mathematics, and thus do not associate mathematics with processes of active exploration. If students do not associate what they have learned in mathematics classes with these processes, they are not likely to apply what they have learned in class to engage in active data exploration inside and outside the classroom. Consequently, there is a need to provide opportunities for students to experience mathematical activities as processes of active inquiry that can also help them to develop data-related skills. Against this background, the present chapter introduces mathematical modeling with hands-on experimentation as one potentially promising approach to engage students in authentic modeling with real data and, thereby, to promote (more) adequate conceptions about mathematics as well as data literacy.

Mathematical modeling with real data gained through hands-on experimentation

Consistent with the aforementioned definition of data literacy by Wolff et al. (2015), mathematical modeling provides students with the opportunity to deal with a real-world problem. Specifically, mathematical modeling refers to "the entire process leading from the original real problem situation to a mathematical model" (Blum & Niss, 1991, p. 39). According to, for example, Blum and Niss (1991) as well as the modeling cycle developed by Blum and Leiß (2007), *mathematical modeling* starts with a real problem situation which first has to be understood, simplified, and structured. This leads to a real model of the original situation that still contains essential features of the original situation, but is also schematized to an extent that it allows a mathematically driven approach. The real model has then to be mathematized by, for instance, translating its data and relations into mathematics, resulting in a mathematical model of the real situation. The mathematical model essentially consists of certain mathematical objects and their relations, which correspond to the core elements and the interaction of these elements of the original real situation or the real model. The mathematical model then allows one to, for instance, draw conclusions, make calculations, apply known mathematical methods, and finally obtain certain mathematical results. These mathematical results have to be re-translated into the real world by interpreting them in relation to the original real situation. The model is thereby also validated, which means that the appropriateness of the results is checked against the background of the real problem situation.

In summary, mathematical modeling enables students to deal with a real-world problem and to move between reality and mathematics. However, when comparing mathematical and statistical modeling, it is apparent that statistical models not only have a deterministic component which may be represented by a mathematical function, but also a stochastic component which provides information about how actual observations deviate from the deterministic component (e.g., Dvir & Ben-Zvi, 2023; chapter by Gould **in this volume on page 81**). Therefore, it seems to be necessary to engage students in mathematical modeling with real data that also includes a stochastic component, and thus may foster students' data literacy. Dealing with real, authentic data can engage students in a broader range of science practices and improve their critical thinking (e.g., Kerlin et al., 2010; Holmes et al., 2015), especially through analyzing and interpreting data, using mathematics and computational thinking, and reasoning based on evidence (e.g., National Research Council, 2012). In contrast, inauthentic data, such as simplified textbook data without noise, are often generated to demonstrate a particular pattern or result from manipulation of data to force a specific result or interpretation (e.g., Kjelvik & Schultheis, 2019) and, thus, often already fit an intended model (e.g., Engel, 2010).

One promising approach to integrate real data in the modeling process, which is focused on in this chapter, is to combine mathematical modeling tasks with data that students gather through mathematical hands-on experimentation (see, e.g., Geisler, 2021a, 2021b; Zell & Beckmann, 2009). In contrast to inner-mathematical experimentation, which asks students to experiment with mathematical objects in the world of symbols (e.g., to examine divisibility rules), *mathematical hands-on experimentation* provides students with a real-world question which they have to investigate through a physical experiment with real objects. They then use mathematics to analyze and evaluate their observations (e.g., Barzel et al., 2007). Thus, mathematical hands-on experimentation can serve as a suitable starting point for mathematical modeling because "experiments related to mathematics find their natural place in the framework of modeling [...] [as] they represent the 'rest of the world' for which mathematical models are built" (Halverscheid, 2008, p. 226). Moreover, as every experiment contains idealizations, the experiment itself can be considered a real model (Geisler, 2021b). During mathematical modeling with hands-on experimentation, students have to make assumptions, plan and conduct an experiment, excerpt a mathematical model from the real world (e.g., by noting the measured values and transferring them to a coordinate system), answer mathematical questions within this mathematical model, interpret the mathematical results in the real situation, validate the solution, and reflect on their approach (see the integrated model of modeling with experiments by Geiser, 2021b; see also Figure 1).

The steps of mathematical modeling with hands-on experimentation are comparable to the approach used in statistical modeling. Specifically, the steps of mathematical modeling with hands-on experimentation correspond to the PPDAC cycle which describes how to abstract and solve a statistical problem grounded in a larger real problem (e.g., Wild & Pfannkuch, 1999). Following the PPDAC cycle, the problem is first understood and defined (Problem). Then, among other things, the measurement system and sampling design are planned (Plan). Afterwards, data are collected, managed, and cleaned (Data). Subsequently, data are explored, and planned as well as unplanned analyses are performed (Analysis). Interpretation, conclusions, and new ideas for future analysis follow (Conclusions).

Through the integration of hands-on experimentation into mathematical modeling, students connect the data they gather, analyze, and interpret to their everyday lives.

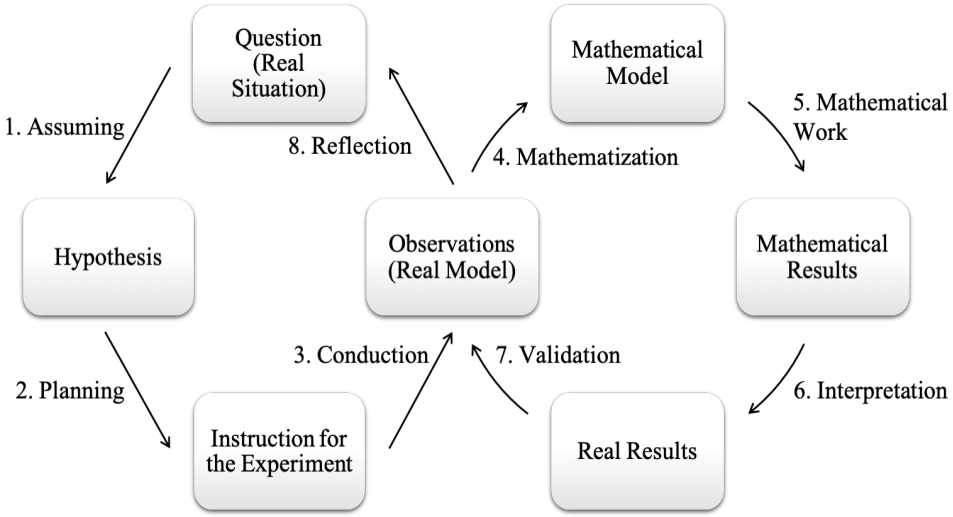


Figure 1. Integrated model of modeling with experiments (adapted from Geisler, 2021b, p. 205)

Thus, it becomes likely to demonstrate the real-world relevance of learning activities focusing on data literacy to students and to trigger their natural curiosity about their world (e.g., Doering & Veletsianos, 2007). Furthermore, when students collect data on their own, they gain a deeper understanding of the data-collection process and its possible limitations (e.g., Hug & McNeill, 2008) and, thus, can take this into account when making their interpretations (e.g., Roth, 1996). Against this background, Kjelvik and Schultheis (2019), for example, argue that students should be given the opportunity to engage in quantitative reasoning and data science while exploring authentic data sets from scientific research in order to gain strong learning experiences surrounding data literacy.

However, the use of complex authentic data, for instance, in terms of scope, selection, curation, size, or messiness, can be challenging for students, especially for novices who have little experience in data exploration or research more generally. For example, Pols et al. (2021) investigated students' ability to analyze experimental data in secondary physics education. Their results indicate that students' performance in interpreting data in terms of the investigated phenomenon or situation was weak. They observed that students were often unable to identify the crucial features of a given graph and conclusions based on the data were often tautological or superficial. In addition, they found that students were not able to infer implications from the data, to interpret data at a higher level of abstraction, or to specify limitations on the validity of the analysis or conclusion. Similar findings were also found in a study by Kanari and Millar (2004) on students' reasoning from data. Against this background, Kjelvik and Schultheis (2019) recommend that students should not start with the exploration of complex data and that the use of simple data sets constitutes a more appropriate starting point. Simple data sets are characterized, for example, by a narrow scope (i.e., are limited to appropriate data) and a small size (i.e., can be explored using pencil and paper and contain few variables and data points). In addition to the potential to promote students' data literacy, using real data for mathematical modeling may also reinforce the importance of mathematics for answering questions and foster students' interest and active engagement in both mathematics and science (e.g., Schultheis & Kjelvik, 2015; Šorgo, 2010).

In summary, mathematical modeling with real data gained through hands-on experimentation seems to be a promising approach to promote students' data literacy, as students are asked to answer a real-world question by using real data they gained through an inquiry process, which corresponds to the definition of data literacy by Wolff et al. (2015). Furthermore, mathematical modeling with hands-on experimentation may also foster (more) adequate conceptions about mathematics in students. However, in light of the findings on students' difficulties in analyzing and interpreting data in science education (e.g., Kanari & Millar, 2004; Pols et al., 2021), it is unclear whether students are able to successfully use real data from mathematical hands-on experimentation to model real-world phenomena mathematically. Therefore, the present chapter provides insights into students' mathematical modeling with real data from hands-on experimentation. Before presenting these findings, we briefly describe the research setting of our study that led to these results.

Research setting

Our data collection took place during a day-long project on mathematical growth and decay processes in an out-of-school lab at a large German university as part of a quasi-experimental field study. Participants were 74 ninth and tenth graders (M_{age} = 14.81, SD = 0.80, 55% male, 42% female, 3% divers) from three German secondary school classes who visited the out-of-school lab as whole classes with their mathematics teachers to attend our day-long project. During the project, the students worked on different mathematical modeling tasks with hands-on experimentation to investigate and model various everyday growth and decay processes.

In the present chapter, we focus on the first learning phase of the day-long project, in which the students had a total of 55 minutes to work in groups of three on a mathematical modeling task with hands-on experimentation on beer foam decay. This context is likely to be motivating and familiar to students from their everyday lives, as beer is interesting and has a positive connotation for many students of this age in Germany (e.g., Wilhelm & Ossau, 2009). Furthermore, this context fulfills the recommendations of Zell and Beckmann (2009, p. 2218f.) for using hands-on experimentation in mathematics lessons. According to their first recommendation, the experiments should be simple—done with few materials and performed quickly—so that students can concentrate more on the mathematics. Moreover, the physical terms used during experimentation should be familiar to the students. As the students were asked to conduct an experiment by measuring the height of foam from poured alcohol-free beer, the experiment was done with few materials (i.e., alcohol-free beer, a measuring cylinder, a ruler, and a stopwatch), took only a few minutes, and the physical terms used (i.e., time and height) were familiar to the students. Because the participating students were not yet familiar with exponential processes, they were not asked to model the beer foam decay as a function. Rather, the beer foam experiment served as an exploration of exponential processes.

While working on the beer foam decay task, the students used a printed lab booklet that contained both the assignments for the subsequent activities and space for students' notes. According to the integrated model of modeling with experiments (see Figure 1), the students first had to individually make assumptions about how the beer foam might decay as well as to draw a sketch of the predicted decay and then discuss their hypotheses with their peers. Afterwards, they were asked to plan an experiment together to investigate the beer foam decay as well as to think about what materials they would need to do so. Before carrying out their planned experiment, the students were asked to think about how to note their measurements. Then they carried out their planned experiment, noting their measured values. Subsequently, in order to support the students in analyzing and interpreting their results as well as to facilitate comparison with their initial assumptions and sketches, the students were asked to first consider how to plot their findings into an appropriate coordinate system and then graph their measured values into their coordinate system. The students had to use their graph to analyze and interpret their results as well as to compare their findings with their initial predictions and sketches. Finally, the students reflected together on their procedure. For this purpose, the students were asked to imagine that they would investigate how beer foam decays a second time and consider what they would do differently and whether they would get the same results and on what conditions the results would depend.

To gain insights into students' mathematical modeling in this task, we analyzed the notes that the students took in their lab booklets. For this purpose, we first examined the notes of the 74 students of all eight steps of the process mentioned above (see also Figure 1) to obtain an overview of the different ways the students worked on the task. In light of the findings on students' difficulties in analyzing and interpreting data in science education (e.g., Kanari & Millar, 2004; Pols et al., 2021), we then focused especially on the phases in which the students were asked to analyze and interpret their results and to compare them with their initial assumptions and sketches. In particular, we explored whether students who subjectively confirmed their initial assumptions tended to (mis)interpret their collected data in line with their initial assumptions (see also Hagenkötter et al., 2023). We jointly selected specific examples to illustrate both the possibilities and difficulties of integrating real data gained through mathematical hands-on experimentation into mathematical modeling.

Insights into students' mathematical modeling with real data gained through hands-on experimentation

We found evidence in the students' notes that mathematical modeling with real data gained through hands-on experimentation can enable students to answer a real-world question from a smaller data set through an inquiry process, which Wolff et al. (2017) describe as part of data literacy. One group of students, for example, assumed at the beginning that "the beer foam dissolves linearly" and drew a corresponding sketch (see Figure 2, A). They chose suitable materials to investigate the decay of beer foam and decided to measure the foam height every 20 seconds. Although the students did not use a table, they systematically noted their measured values, albeit in a mathematically questionable form (see Figure 2, B). They plotted them into a suitable coordinate system, albeit without labeling axes, and their graph represented an approximate exponential decay (see Figure 2, C). The students described the beer foam decay using their graph as follows: "The foam decays much faster at the beginning than at the end." Consequently, they concluded: "The foam did not decay as expected. We thought that it would decay evenly, but it decays quickly first and then more and more slowly at the end." Finally, the students reflected that "the experiment was good as it was." They correctly did not expect other results, i.e., no other process of decay, to occur when they investigate the beer foam decay a second time, but "there could be discrepancies in the measurements, of course." Moreover, one student of the group referred, albeit in a more general and less specific way, to possible conditions that might have influenced the process of decay: "I do not think that other results will occur, because we are in the same room with the same conditions."

However, we also observed that dealing with real data gained through mathematical hands-on experimentation can be challenging for students. With regard to the students who subjectively confirmed their initial assumptions during mathematical modeling with hands-on experimentation (see also Hagenkötter et al., 2023), our results show that many students tended to interpret their results in a biased way and, consequently, falsely confirmed their initial assumptions. One group, for example, incorrectly confirmed their initial assumption by focusing their interpretation on very general aspects of their results that were consistent with their initial assumption. The students initially assumed that "the foam decreases proportionally" and drew a corresponding sketch (see Figure 3, A). They also chose suitable materials to investigate the decay of beer foam and decided to measure the foam height every five seconds. The students again did not use a table, but they systematically noted their measured values (see Figure 3, B). They plotted them into a coordinate system, but reversed the independent and dependent variables (see Figure 3, C). Nevertheless, the results of their experiment clearly showed no linear decay, but instead an approximately exponential decay. Although the students were asked to describe the beer foam decay using their graph, the students did not describe the actually observed beer foam decay, but only their graph in a very general way as follows: "We have noticed that the graph has decreased." The students did not describe the decay process in more detail and, thus, ignored that the decay process did not correspond with their initial assumption of a linear decay. Finally, they erroneously concluded that the results of their experiment support their initial assumption: "assumption confirmed through experiment!"

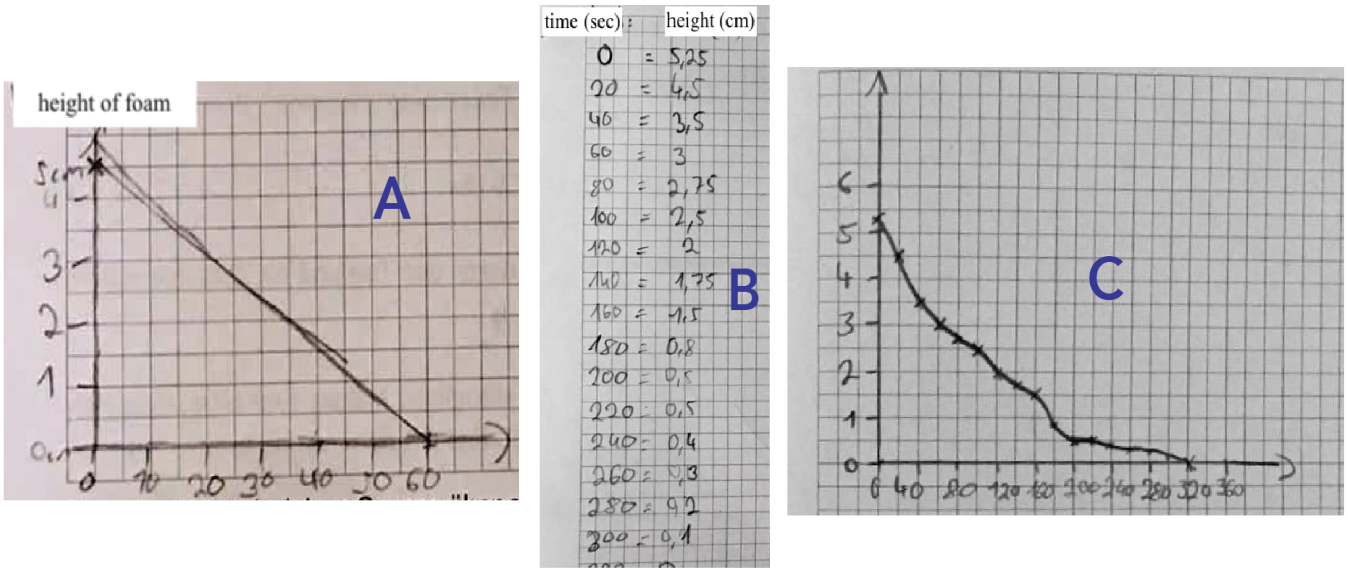


Figure 2. Sketch of the initial assumption that beer foam dissolves linearly (A), measured values of the students (B), and graph with the collected data that represents an approximate exponential decay (C)

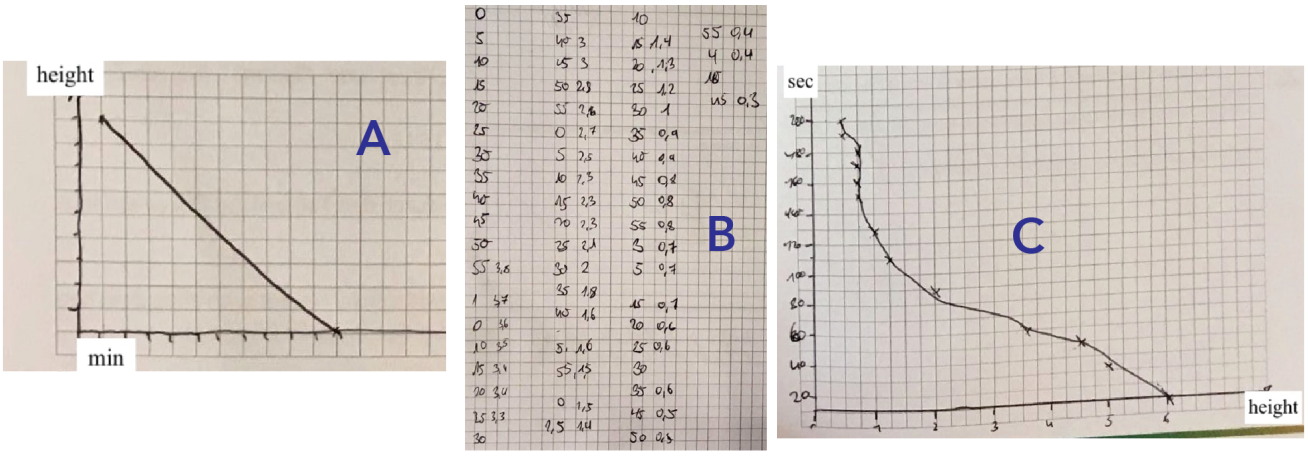


Figure 3. Sketch of the initial assumption that beer foam decreases proportionally (A), measured values of the students (B), and graph with the collected data that represents an approximate exponential decay (C)

Discussion and conclusion

During mathematical modeling with hands-on experimentation, students usually do not work with large data sets, but they nevertheless answer real-world questions using real data that they collected on their own. By identifying, collecting, analyzing, interpreting, and evaluating data, students are taking first steps towards becoming active data explorers (see Oceans of Data Institute, 2015) and, thus, get the opportunity to develop data literacy competencies (see, e.g., Wolff et al., 2015). As the findings we presented in this chapter show, students can successfully use their collected data to mathematically model real-world phenomena such as the decay of beer foam. Moreover, collecting their own data enabled students to reflect more strongly on the circumstances when interpreting the data. In addition, mathematical modeling with hands-on experimentation has the potential to reinforce the importance of mathematics for answering questions and, thus, to promote (more) adequate conceptions about mathematics like the following quote of a student after working on various mathematical modeling tasks with hands-on experimentation shows: "It was easy to recognize and understand the connection between math and reality."

However, although we used a simple data set, our results show that the use of real data can be challenging for students (see, Kjelvik & Schultheis, 2019). Besides typical mistakes of students in dealing with functions, such as reversing the dependent and independent variable (see, e.g., Hofmann & Roth, 2021), we observed that the students often had problems interpreting their graphs in light of the real-world situation. As demonstrated by the second example we provided, the students planned and carried out a suitable experiment and first systematically noted and then transferred their measured values to a coordinate system, but they provided a naïve interpretation of their experimental results. We explicitly asked the students to compare their findings with their initial assumptions and sketches, which has been shown to support students' dealing with statistical models and modeling in the context of informal statistical inference (e.g., Dvir & Ben-Zvi, 2018). Nevertheless, the example presented and additional solutions by other students indicate that the students often described the decay process in very general terms, tending to confirm their initial assumptions. Thus, our findings are in line with the results of previous studies on students' difficulties in analyzing and interpreting data in science education (e.g., Kanari & Millar, 2004; Pols et al., 2021), which also indicate that students' conclusions based on data are often tautological or superficial. Based on the discussions about the unique affordances of digital data visualization and modeling tools at Minerva School, it seems promising to additionally support students through the use of tools such as CODAP (Finzer, 2019) or TinkerPlots (Konold & Miller, 2005). This may also enable students to first collect and interpret their own data set and then combine their data with other students' data sets as well as with larger, professionally collected data sets that encompass and extend beyond the circumstances of their self-collected data (see, e.g., nested data set strategy by Kastens et al., 2015, p. 28).

To conclude, our insights and, in particular, students' difficulties in mathematical modeling with hands-on experimentation suggest that students first need to learn to mathematically model real-world phenomena and, especially, to deal with real data. Therefore, it seems necessary to explicitly discuss with students how to deal with real data and practice it repeatedly from the beginning of their school education. Even though mathematics classes are particularly suitable for fostering data literacy in students, the use of real data can and should also be promoted in other subjects and contexts such as biology

(see, e.g., Bar, 2022), computer science (see, e.g., the chapter by Podworny & Frischemeier **in this volume on page 15**), or machine learning (see, e.g., Fleischer, 2022). Furthermore, it seems useful to teach data literacy together with other closely related competence areas, such as computational thinking or logical reasoning (see, e.g., Labusch, 2022), or in an interdisciplinary way in order to contextualize the concepts and methods and, thereby, to promote transferable skills

References

Bar, C. (2022). Dataset-driven instruction in the biology classroom. Paper presented at the Minerva School 2022: *Reasoning with data models and modeling in the big data era.*

Barzel, B., Büchter, A., & Leuders, T. (2007). Experimentieren [Experimentation]. In B. Barzel, A. Büchter, & T. Leuders (Eds.), *Mathematik-Methodik.* (1. Ed., pp. 70–75). Berlin: Cornelsen Scriptor.

Blum, W. & Leiß, D. (2007). Ho do students and teachers deal with modelling problems? In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical Modelling. Education, Engineering and Economics—ICTMA 12* (pp. 222–231). Chichester: Horwood Publisihing. https://doi.org/10.1533/9780857099419.5.221

Blum, W. & Niss, M. (1991). Applied mathematical problem solving, modelling, applications, and links to other subjects—State, trends and issues in mathematics instruction. *Educational Studies in Mathematics, 22*, 37–68. https://doi.org/10.1007/BF00302716

Doering, A., & Veletsianos, G. (2007). An investigation of the use of real-time, authentic geospatial data in the K–12 classroom. *Journal of Geography, 106*(6), 217–225. https://doi.org/10.1080/00221340701845219

Dvir, M. & Ben-Zvi, D. (2018). The role of model comparison in young learners' reasoning with statistical models and modeling. *ZDM Mathematics Education, 50*, 1183–1196. https://doi.org/10.1007/s11858-018-0987-4

Dvir, M. & Ben-Zvi, D. (2023). Informal statistical models and modeling. *Mathematical Thinking and Learning, 25*(1), 79–99. https://doi.org/10.1080/10986065.2021.1925842

Engel, J. (2010). *Anwendungsorientierte Mathematik: Von Daten zur Funktion* [Applied Mathematics: From Data to the Function]. Berlin: Springer Verlag.

Finzer, W. (2019). *Common Online Data Analysis Platform (CODAP).* Concord: The Concord Consortium.

Fleischer, Y. (2022). Reasoning with predictive models—teaching about machine learning with decision tree as approach to predictive modeling. Paper presented at the Minerva School 2022: *Reasoning with data models and modeling in the big data era.*

Geisler, S. (2021a). Data-based modelling with experiments—Students' experiences with model-validation. In M. Inprasitha, N. Changsri & N. Boonsena (Eds.), *Proceedings of the 44th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, p. 330–339). Khon Kaen: PME.

Geisler, S. (2021b). Mathematical modelling with experiments—Suggestion for an integrated model. In M. Inprasitha, N. Changsri & N. Boonsena (Eds.), *Proceedings of the 44th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 1, p. 205). Khon Kaen: PME.

Gould, R. (2024). **"Traditional statistical models in a sea of data: teaching introductory data science" in this volume on page 81.**

Hagenkötter, R., Nachtigall, V., Rolka, K., & Rummel, N. (2022). Exploring students' and mathematics teachers' conceptions about the work of mathematical scientists and possible relations to mathematics teaching. In C. Chinn, E. Tan, C. Chan & Y. Kali (Eds.), *Proceedings of the 16th International Conference of the Learning Sciences—ICLS 2022* (pp. 155–162). Hiroshima: ISLS.

Hagenkötter, R., Nachtigall, V., Rolka, K., & Rummel, N. (2023). "Our result corresponds with our assumption"—Students' biased interpretation of data gained through mathematical hands-on experimentation. In P. Blikstein, J. van Aalst, R. Kizito, & K. Brennan (Eds.), *Proceedings of the 17th International Conference of the Learning Sciences—ICLS 2023* (pp. 1901–1902). Montreal: ISLS.

Halverscheid, S. (2008). Building a local conceptual framework for epistemic actions in a modelling environment with experiments. *ZDM Mathematics Education, 40*, 225–234. https://doi.org/10.1007/s11858-008-0088-x

Hofmann, R. & Roth, J. (2021). Lernfortschritte identifizieren. Typische Fehler im Umgang mit Funktionen [Identifying learning progressions. Typical mistakes when dealing with functions]. *mathematik lehren, 226*, 15–19.

Holmes, N. G., Wieman, C. E., & Bonn, D. A. (2015). Teaching critical thinking. *Proceedings of the National Academy of Sciences USA, 112*(36), 11199–11204. https://doi.org/10.1073/pnas.1505329112

Hug, B., & McNeill, K. L. (2008). Use of first-hand and second-hand data in science: Does data type influence classroom conversations? *International Journal of Science Education, 30*(13), 1725–1751. https://doi.org/10.1080/09500690701506945

Kanari, Z. & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Education, 41*(7), 748–769. https://doi.org/10.1002/tea.20020

Kastens, K., Krumhansl, R., & Baker, I. (2015). Thinking big. Transitioning your students from working with small, student-collected data sets towards "big data." *The Science Teacher, 82*(5), 25–31.

Kerlin, S. C., McDonald, S. P., & Kelly, G. J. (2010). Complexity of secondary scientific data sources and students' argumentative discourse. *International Journal of Science Education, 32*(9), 1207–1225. https://doi.org/10.1080/09500690902995632

Kjelvik, M. K., & Schultheis, E. H. (2019). Getting messy with authentic data: Exploring the potential of using data from scientific research to support student data literacy. *CBE Life Sci Educ, 18*(2). https://doi.org/10.1187/cbe.18-02-0023

Konold, C. & Miller, C. D. (2005). *TinkerPlots: Dynamic Data Explorations.* Emeryville: Key Curriculum Press.

Labusch, A. (2022). Eighth graders' data modeling within their computational thinking and problem-solving. Paper presented at the Minerva School 2022: *Reasoning with data models and modeling in the big data era.*

National Research Council (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: National Academies Press.

Oceans of Data Institute (2015). *Building Global Interest in Data Literacy: A Dialogue.* Workshop Report. Waltham: Educational Development Center.

Podworny, S. and Frischemeier, D. (2024). **"Young learners' perspectives on the concept of data as a model: what are data and what are they used for?" in this volume on page 15.**

Pols, C. F. J., Dekkers, P. J. J. M., & de Vries, M. J. (2021). What do they know? Investigating students' ability to analyse experimental data in secondary physics education. *International Journal of Science Education, 43*(2), 274–297. https://doi.org/10.1080/09500693.2020.1865588

Risdale, C., Rothwell, J., Smit, 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.*

Roth, W.-M. (1996). Where is the context in contextual word problems? Mathematical practices and products in grade 8 students' answers to story problems. *Cognition and Instruction, 14*(4), 487–527. https://doi.org/10.1207/s1532690xci1404_3

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Groues (Ed.), *Handbook for research on mathematics teaching and learning* (pp. 334–370). New York: Macmillan.

Schultheis, E. H., & Kjelvik, M. K. (2015). Data nuggets: Bringing real data into the classroom to unearth students' quantitative and inquiry skill. *American Biology Teacher, 77*(1), 19–29.

Šorgo, A. (2010). Connecting biology and mathematics: First prepare the teachers. *CBE—Life Sciences Education, 9*, 196–200. https://doi.org/10.1187/cbe.10-03-0014

Wild, C. J. & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–265. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Wilhelm, T. & Ossau, W. (2009). Bierschaumzerfall—Modelle und Realität im Vergleich [Beer foam decay – Comparison between models and reality]. *Praxis der Naturwissenschaften – Physik in der Schule, 58*(8), 19–26.

Wolff, A., Gooch, D., Cavero Montaner, J. J., Rashid, U., & Kortuem, G. (2015). Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics, 12*(3), 9–26. https://doi.org/10.15353/joci.v12i3.3275

Zell, S. & Beckmann, A. (2009). Modelling activities while doing experiments to discover the concept of variable. In V. Durand-Guerrier, S. Soury-Lavergne, & F. Arzarello (Eds.), *Proceedings of the Sixth Congress of the European Society for Research in Mathematics Education* (pp. 2216–2225). Lyon: INRP.

Design principles for developing statistical literacy by integrating data, models, and context in a digital learning environment

Christian Büscher

University of Duisburg-Essen
christian.buescher@uni-due.de

The development of statistical literacy is a key challenge to teachers in the 21st century, but little guidance for designing learning environments for the development of statistical literacy can be found in literature. This chapter reports on a design research project developing a digital learning environment that implements three design principles: build context knowledge, focus on models as evidence for claims, and elicit and scaffold reflections of argumentation. Central to the learning environment is the investigation of a critical climate change context, the Arctic sea ice decline. Results of 12 design experiments conducted with 24 Grade 5 students provide insights into the mechanisms of the design principles and into students' learning processes for developing statistical literacy.

Introduction

Statistical literacy is a fundamental skill that is required of all citizens in the 21st century (Wild, 2017). Recent years have impressively demonstrated this fact, as global crises of climate change and the COVID pandemic have been matched by an increasing amount of fake news and manipulative posting on social media. For statistics education research, this is not a newly discovered fact, as for years, researchers have highlighted the importance of civic statistics for citizens to engage in public discourse (Engel, this volume; 2017). The media reporting concerning the recent COVID pandemic has inspired researchers to uncover the high critical demands that media items pose on readers, as a solid statistical, mathematical, and critical knowledge base is required to understand and possibly criticize such media items (Gal & Geiger, 2022).

In statistics education research, such demands are discussed under the construct of statistical literacy (Gal, 2002). However, while solid theoretical foundations for statistical literacy exist, only few theoretically and empirically grounded concrete approaches for developing statistical literacy can be found in literature (Büscher, 2022), a finding that has led to calls for researchers to investigate such approaches (Ben-Zvi et al., 2018). A notable recent contribution to this call is provided by Gal (2024, **in this volume on page 91**), who outlined general recommendations for including statistical discussions on societal problems in the teaching of statistics. This chapter continues this work and provides a further specification of concrete activities and tasks by reporting on a design research study that investigates approaches for developing statistical literacy in middle school mathematics classrooms. Design research can provide a useful methodological framework for identifying useful design principles that can inform the design of learning environments. The aim of this chapter is to illustrate the working mechanisms of three identified design principles by investigating the learning processes of students working in a digital learning environment that implements these three design principles. The theoretical framework presents a perspective on statistical literacy that pays attention to the role of models for reflecting on statistical argumentation and provides the theoretical grounds for three identified design principles. The following section outlines a concrete implementation of the design principles to serve as guidance for future designers of learning environments. The empirical section then provides insights into the initiated learning processes and the working mechanisms of the identified design principles.

Theoretical framework

A modeling perspective on critical statistical literacy

Gal (2002) defined statistical literacy as the ability to understand and to evaluate statistical information as well as to communicate one's reactions towards this information. Weiland (2017) proposed to extend the notion of statistical literacy to *critical* statistical literacy to include an important observation about statistical arguments in society:

> "Statistical arguments are not made from an objective independent reality. They are made by individuals from a multitude of subjectivities. In this

> sense statistical arguments can serve to perpetuate discourses." (Weiland, 2017, p. 42).

This means that statistical arguments cannot be evaluated solely based on fit to an objective reality. Instead, statistical arguments are made by individuals who are situated in a social context to participate in a discourse that comprises social, political, economic, ecological, and other factors. A similar point is also made by Gal (2024, **in this volume on page 91**) who calls for statistics education to adopt an "external view" on statistical models that also takes into account the role of models in different contexts such as media contexts, service consumption contexts and workplace contexts. Jablonka and Bergsten (2021) also take an external view on models by illustrating the argumentative strategies in which models are used by policy makers in public discourse, which would not be described sufficiently using an "internal view" of models. These authors show that statistical literacy does not only refer to the need for citizens to be able to understand the diagrams and measures used in the statistical argument at hand, but also to the need to understand the way the model-based argument shapes, and is shaped by, the discourse in which it takes place. Such a critical statistical literacy can then be a tool for challenging the power structures observed in society which are coded into statistical data and arguments, and which are reported through media items or government reports.

Such statistical arguments do not simply list data, but instead build on models to emphasize the relationships in data. To understand the role of models and modeling for critical statistical literacy, it is helpful to draw on ideas proposed by Skovsmose (1994), who builds on a central observation regarding models:

> "Thus, any modeling process presupposes that certain simplifications are established. This means paying attention to certain aspects of 'reality' and neglecting others." (Skovsmose, 1994, p. 199).

For Skovsmose (1998), citizens confronted with a model that pays attention to certain aspects of reality while neglecting others need to be able to engage in *reflection* under four different orientations: mathematics-oriented reflection regarding the correctness of calculations and formal characteristics; model-oriented reflection regarding the fit of an employed model to a larger phenomenon; context-oriented reflection on the possible consequences the use of mathematics has on the phenomenon; and lifeworld-oriented reflection relating to the subjective meaning of mathematics for the reflecting citizen.

To highlight the role of data and models for reflection, this study adapts the four ideas relating different orientations of reflection for critical statistical literacy using a theoretical model that conceptualizes the demands of critical statistical literacy for an individual confronted with a statistical argument (Figure 1, adapted from Büscher, 2022). In this theoretical model, statistical arguments consist of elements taken from four different *domains of argumentation*:

- *Phenomena* about which the arguments aim to establish certain claims or to justify certain actions;
- *Data* which provide the quantitative basis of the claim about the phenomenon;
- *Models* that abstract from the data to describe a specific relationship found in the data; and
- *Text* that provides illustrations and is composed in a certain style of rhetoric. 'Text' is here taken in a general way, meaning not only written words but also illustrations, graphs, and other forms of representation.

This conceptualization builds on a broad understanding of models as "purposeful mathematical descriptions of situations, embedded within particular systems of practice that feature an epistemology of model fit and revision" (Lesh & Lehrer, 2003, p. 109). They are distinguished from the data in that data provide encoded representations of the "facts" of situations, while models represent the mathematical relationships that can be found within the data.

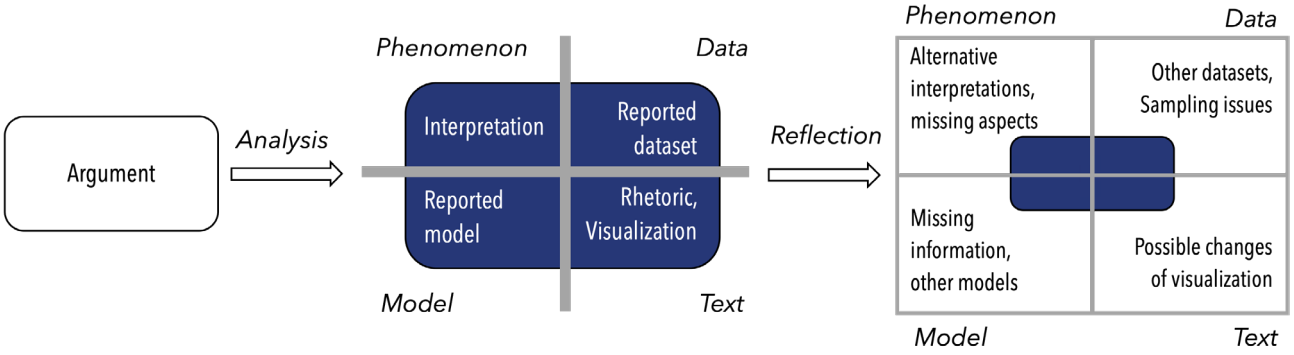


Figure1 Critical statistical literacy as analysis and reflection of statistical arguments in four domains of argumentation

In argumentation, these domains are always interconnected: phenomena are described via data, data form the background of models, models are illustrated through text, and text provides information on phenomena, data, and models. Yet each domain of argumentation comes with its own set of criteria for evaluating a statistical argument. Confronted with an argument, a reader or listener has to engage in *analysis* to disentangle the web of interconnected domains of argumentation. An analysis in the domain of the *phenomenon* can consist of asking questions like: what natural or social phenomenon is the argument about? Which aspect of this phenomenon does the argument focus on here? What is the interpretation of the results provided in the argument? In contrast, an analysis in the domain of the *data* asks: how were the data obtained? What types of variables are given? What were the methods of sampling? Building on this, an analysis in the domain of the *model* can be structured along questions like: Which statistical indicators were used? What type of regression was applied? And finally, an analysis in the domain of the *text* asks questions concerning the presentation of the other domains: what kind of rhetoric is used? How is the model illustrated? Which arguments are made explicit?

Because models are purposeful representations of situations, they are not neutral, but encode a certain subjectivity (Weiland, 2017). Critical statistical literacy highlights the need for not only analyzing and understanding an argument, but also to uncover any possibly hidden agenda that influenced its creation. Therefore, a reader needs to engage in *reflection*. Instead of organizing different types of reflections along different orientations (Skovsmose, 1998), again the domains of argumentation are used here. A reflection in the domain of the phenomenon might ask: What is known about the phenomenon that does not appear in the argument? What might be the agenda of someone arguing in this way? A reflection in the domain of the data might consist of questions like: which data were not taken into account? Could other sampling methods have produced other data? A reflection in the domain of the model might ask: would a different type of regression or a different measure of center have produced a different result? Was there any information missing that would have been important for interpreting the model? And finally, a reflection in the domain of the text asks questions such as: would a different visualization lead to a different interpretation? Would a different rhetoric be more appropriate? This specification of reflection highlights that the model-oriented reflection proposed by Skovsmose (1998) touches on very different domains of argumentation, as the question of fit between model and situation needs to touch on alternatives regarding phenomenon, data, and model. Thus, it shows that models play a central role in reflection about arguments.

In this way, analysis and reflection engage with argumentation using the same domains of argumentation, but in different ways: the analysis of an argument identifies and explicates the internal information used in the argument along the different domains of argumentation, whereas reflection uncovers missing information that is external to the argument at hand. This shows the high demands faced by learners in developing critical statistical literacy. Equally, it shows the high demands faced by educators for supporting learners in navigating this complex interplay of internal and external relations between phenomena, data, model, and text. One approach to help educators is to specify design principles that can inform the design of learning environments and teaching units, which will be presented in the following section.

Design principles for developing critical statistical literacy

Design principles that provide guidance for future designers of learning environments are one of the major outcomes of design research (van den Akker, 1999; Prediger et al., 2015). While elaborated design principles arise from empirical design and research work, they also need to be grounded in theory (van den Akker, 1999). Although there are not many explicit design principles mentioned in the literature on statistical literacy (Büscher, 2022), some general advice on how to construct a learning environment fostering statistical literacy can still be found throughout various studies. For example, Gal and colleagues (2023) recommended promoting engagement with societal issues, using socially relevant data and text, using technologies that enable rich visualizations and developing skills of critical interpretation. In the following, these ideas are subsumed under three core design principles which build on the theoretical specification sketched earlier.Build critical context knowledge. In his conceptualization of statistical literacy, Gal (2002) already specified context knowledge as a central knowledge element of statistical literacy. Weiland (2017) observed that for many critical statistical literacy tasks, critical contexts surrounding issues of race, gender, or climate change need to be addressed in classrooms. This presents a challenge to many educators, as such contexts are often not part of teacher training, and teachers might not feel comfortable integrating such issues into their teaching. The role of context knowledge was also observed by Stephan and colleagues (2021), who found that it is easier for students to show a critical consciousness of mathematics in contexts that they have direct experiences with. This shows that in order to develop critical statistical literacy, learning environments not only have to use data from critical contexts, but also have to actively build the students' context knowledge. This could mean providing not only facts but also

opinions, viewpoints, and ideas surrounding a phenomenon, so that the data can be embedded in a rich network of contextual meaning.

Focus on models as evidence for claims. The role of evidence for argumentation has long been discussed in statistics education research, for example regarding the constructs of Informal Statistical Inference (Makar & Rubin, 2009) and Informal Inferential Reasoning (Zieffler et al., 2008). Makar and Rubin (2009) considered using data as evidence as a central pillar of Informal Statistical Inference and found that young students struggle to see data as useful evidence for their claims. Regarding Informal Inferential Reasoning, the notion of evidence is closely linked to the concept of sampling (Zieffler et al., 2008). Here, stronger evidence is conceptualized as arising mostly from larger or better samples. While these considerations surrounding data as evidence are important, the modeling perspective on statistical literacy employed in this chapter highlights a different perspective on evidence. Since models are abstracted from data and show only selected relationships within the data, different models can be used as evidence for different claims, even within the same dataset. The question whether a claim is backed by evidence cannot be answered solely on the basis of how the data were sampled. Evidence for a claim needs to refer to a specific relationship found in the data i.e., a model, and the act of constructing a model is always a creative act of selecting the subjectively relevant aspects to be modeled, informed by context knowledge. Thus, a learning environment for developing statistical literacy should provide learning opportunities for students to understand the way models act as mediators between data and claim about a phenomenon.

Elicit and support reflection on argumentation. Most design principles found in statistics education research are focused on engaging students in rich inquiry activities and data exploration. For example, the framework of the Statistical Reasoning Learning Environment (Garfield & Ben-Zvi, 2008) outlines a learning environment in which students use technological tools in activities eliciting statistical reasoning in real and motivating contexts. The goal is to develop students' central statistical concepts like distribution or variability. However, the goal of a learning environment for developing statistical literacy is not to develop statistical concepts like center, but to develop insights into the role of models in producing evidence, which requires a different kind of knowledge and different type of learning activity. This point is made by Büscher and Prediger (2019), who conceptualized insights into the nature of models as "reflective concepts" like the concept of perspectivity: the idea that models always create a certain perspective on a phenomenon which highlights specific aspects while simultaneously obscuring others. They proposed that the development of reflective concepts requires students to engage in "reflective activities" like explicating the uses and limits of a statistical measure. Regarding critical statistical literacy, this means that students need to engage in activities of reflecting on argumentation.

These three design principles might serve as general guidance for designing learning environments for developing critical statistical literacy. Yet research and practice require more than general outlines, and recent calls have surfaced for researchers to report on the development of learning environments (Ben-Zvi et al., 2018). Design principles are not only theoretical considerations, but their working mechanisms need to be elaborated by empirically identifying their effects. The remainder of this chapter therefore empirically investigates the three design principles by engaging with the following research question:

Research question: How do the three design principles support the development of students' ability to analyze and reflect on statistical arguments?

Methodological framework

This design research study was carried out with a focus on students' learning processes (Prediger et al., 2015). The overarching *cli.math* project aims to develop and to investigate a digital learning environment that employs climate contexts for developing students' critical statistical literacy and to uncover typical learning pathways and possible obstacles.

Participants and data collection

In June 2022, 12 design experiments (Gravemeijer & Cobb, 2006) were carried out. Each design experiment consisted of one pair of students who worked with the cli.math digital learning environment (see below) using a supplied laptop under supervision of the author, who acted as interviewer and teacher during the experiments. All 24 students were taken from the same mathematics class, a German Grade 5 mathematics class from a school in a German low-income metropolitan area. The students all volunteered for the design experiments, and all students who volunteered were included in the experiments. Each design experiment lasted for about 40 minutes. Video data was captured from each whole design experiment, resulting in approximately 480 minutes of video data. The design experiments are subject to an ongoing analysis of transcribed video data combining deductive and inductive steps of analysis and category generation in the style of open and axial coding (Corbin & Strauss, 1990). The analysis reconstructed the students' references to phenomena, data, model, and text and identified them as internal (i.e., referencing an element of the learning environment) or external (i.e.,

referencing an element taken from a different source such as subjective knowledge). This made it possible to identify the students' processes of analyzing or reflecting on arguments which were either provided by the learning environment or constructed by the students themselves. Afterwards, these references were compared to the design principles to illustrate the effect the design principles had on the students' learning processes.

The cli.math digital learning environment

The cli.math digital learning environment is a browser-based learning environment that was fully coded by the author using *php* and *javascript* as well as the *chart.js* and *konva* libraries. In the learning environment, students' progress through three "worlds": (1) the *story world*, which focuses on the design principle of building context knowledge and in which students discover varied information and claims about a phenomenon; (2) the *data world*, in which students engage in statistical investigation themselves and which focuses on the design principles of using models as evidence for claims; and (3) the *argument world*, which focuses on the design principle of eliciting and scaffolding reflection of argumentation by letting students evaluate claims and evidence produced by others.

The story world

The story world is the first world students encounter in the cli.math digital learning environment (Figure 2). The goal of this stage of the activity is to build a context knowledge base. The central design element here is the mechanism of *collecting info cards and claim cards*. Info cards represent known, undisputed facts about the phenomenon, while claim cards represent claims that can be investigated via the data provided by the learning environment (Figure 3). Although the information on the info cards could also be disputed, they are presented as facts in order to provide the students with a firm basis for interpreting the data which is later provided to them. For example, one info card states that the Arctic sea ice is naturally influenced by summer and winter months. This is a fact that is important to adequately interpret Arctic sea ice data, yet is not a well-known fact to students in Grade 5.

The cards already collected by the students are placed in the card bar at the top of the page (Figure 2, ❶). At any time, the students can click on a card image to view the collected card. To collect new cards, the students engage with different *stories* (Figure 2, ❷). These stories are first presented via a thumbnail picture, which the students can choose to investigate. In this implementation, stories come in text form, and provide articles (here: a fictional wiki-like article on the Arctic), interviews (here: a fictional interview with a polar researcher about Arctic sea ice during summer and winter), or tweet-like social media posts (here: fictional tweets about Arctic sea-ice decline as well as misinformation or debatable claims). Video and audio entries are planned, but not yet implemented.

In this example, a tweet is being investigated that claims that Arctic sea ice will rise again and thus there is no need to worry (Figure 2, ❸). Although this is a fictitious example, similar tweets are easily found in social media, and students are familiar with the need not to take all social media posts at face value. After reading this tweet,

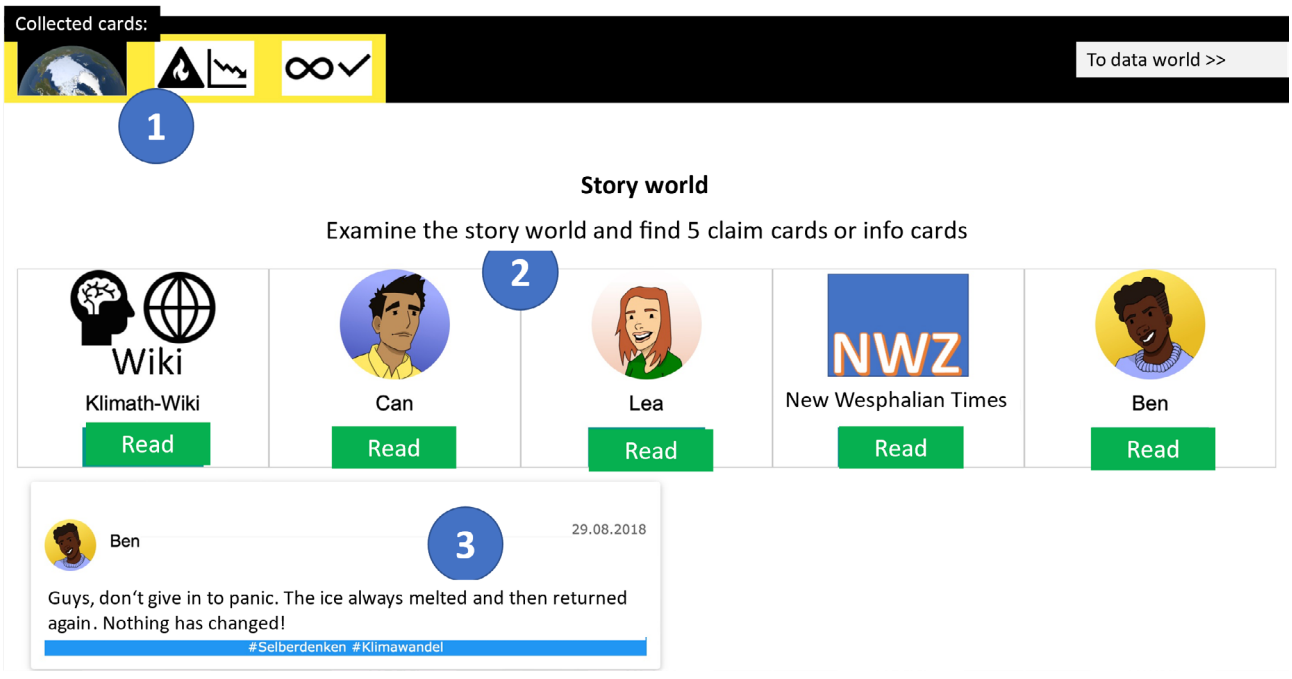


Figure 2: The cli.math story world

students can collect a claim card about not needing to worry, which like all claim cards is marked as in need of verification. After collecting all 5 cards, the students are asked to continue to the data world.

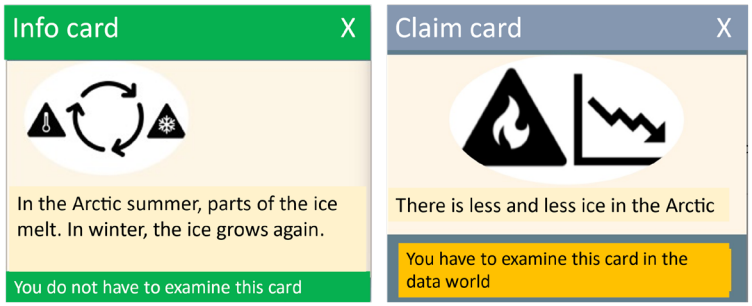


Figure 3: Info card (left) and a claim card (right) provide the context knowledge base of the learning environment

The data world

In the data world (Figure 4), students investigate the claim cards collected throughout the story world. In the card bar, the students can view info cards and choose a claim card to be investigated (Figure 4, ❶). For this, they can choose from different datasets (Figure 4, ❷), not all of which are useful for investigating the claim card. In this example, the students can choose between data on Arctic temperatures and Arctic sea ice extent in two different time frames. Here, the students have chosen the dataset on Arctic sea ice extent during the five "decade" years from 1980 to 2020. After choosing a dataset, the students can view the data through different representations: a table, a bar chart and a line graph (Figure 4, ❸), with the table view being the default representation that is displayed after choosing a dataset. Here, the students have opted to view the line chart (Figure 4, ❹). In the next step, the students have various tools for dealing with the graph (Figure 4, ❺). One important tool here is the box tool for highlighting areas of importance within the data (see Figure 5). Other possibilities include the calculation of statistical measures like mean, mode, and range. More complex representations like dynamic hat plots are planned, but not yet implemented.

When the students are satisfied with their representation, they can take a *data snapshot* (Figure 4, ❻). This creation of data snapshots is the central design element of the data world. Here, they connect the claim card with their representation and are asked to provide an explanation (Figure 5). In this case, they have chosen the box tool and have created a red box highlighting the summer months in the graph. In this way, the students are scaffolded in creating arguments that connect phenomenon, data, and model. Afterwards, the students can progress to the argument world.

The argument world

The argument world focuses on the design principle of eliciting and scaffolding reflection on arguments (Figure 6). Here, the students can choose between different data snapshots created by fictitious students (Figure 6, ❶). In this case, the argument of "Mei" has been chosen, and a claim card, an annotated bar chart and an argument is provided (Figure 6, ❷). This argument has been constructed in a way that it is correct, given this chosen data. To criticize this argument, students have to draw on knowledge about the phenomenon, data, or models external to the argument at hand, that is, they have to *reflect* on the given argument. This reflection is scaffolded by criteria that are provided to the students which they are asked to use to rate Mei's argument (Figure 6, ❸). The exact formulations of the criteria are tentative, and

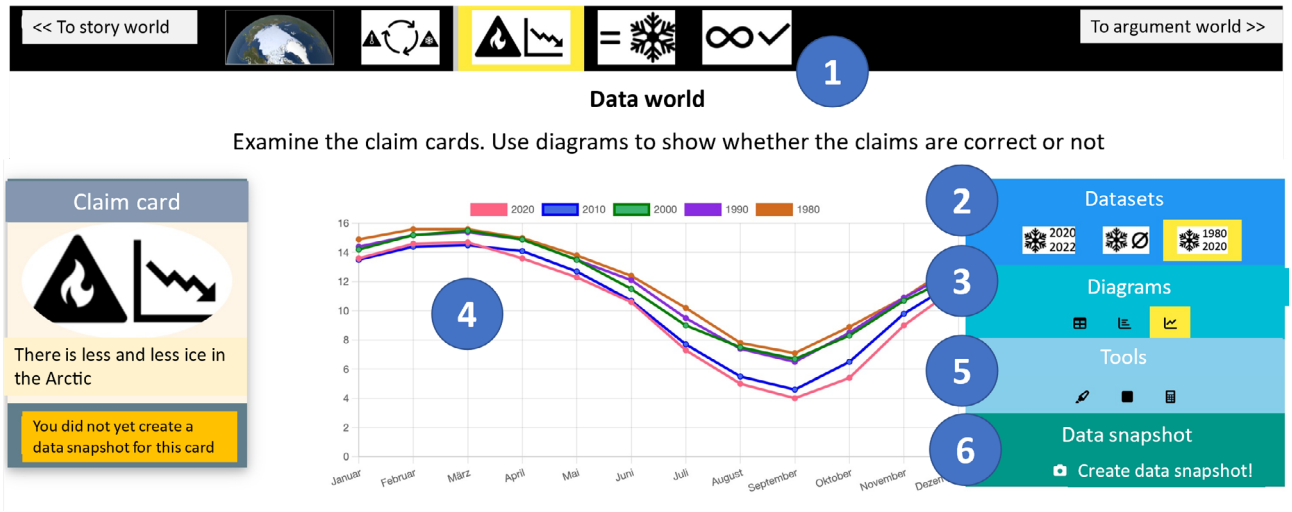


Figure 4: The cli.math data world

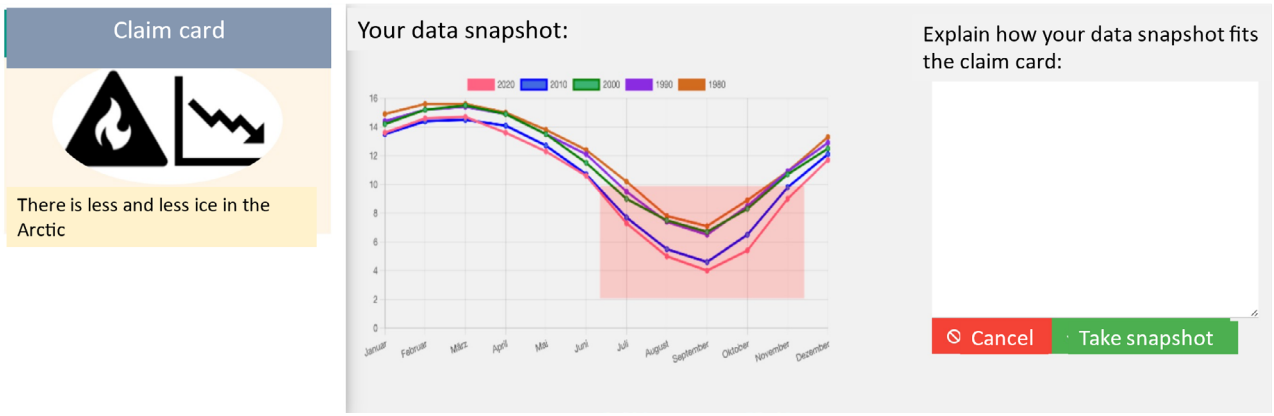


Figure 5: Creating a data snapshot

subject to change during the design research project. In this iteration, criteria like "argument and data snapshot are fair" are supposed to initiate reflections about data and models. Building on these ratings, the students are asked to provide a final opinion on the provided argumentation (Figure 6, ❹). Here, they have the option to use provided sentence fragments (Figure 6, ❺; Figure 7) as scaffolds for formulating a critique of an argument. As before, these scaffolds are a first attempt at providing useful scaffolds, and the design research project investigates whether, and how, such scaffolds can be useful.

Empirical insights

The empirical part of this chapter illustrates the mechanisms of the design principles by showing the activities of one pair of students, Cedrik (C) and Dominik (D), during the whole experiment. The pair was chosen as they worked on all worlds in the learning environment and showed rich engagement with the design principles, although not all pairs profited as much from the design principles as Cedrik and Dominik.

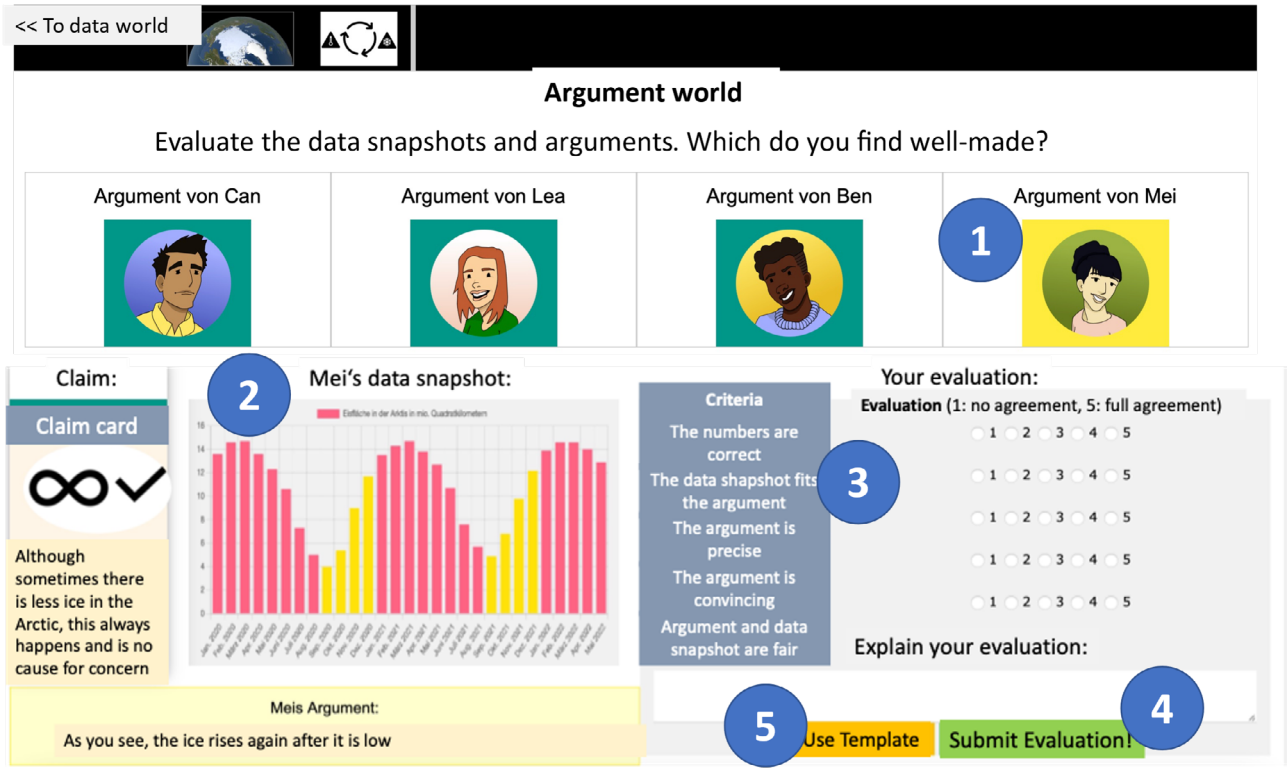


Figure 6: The cli.math argument world

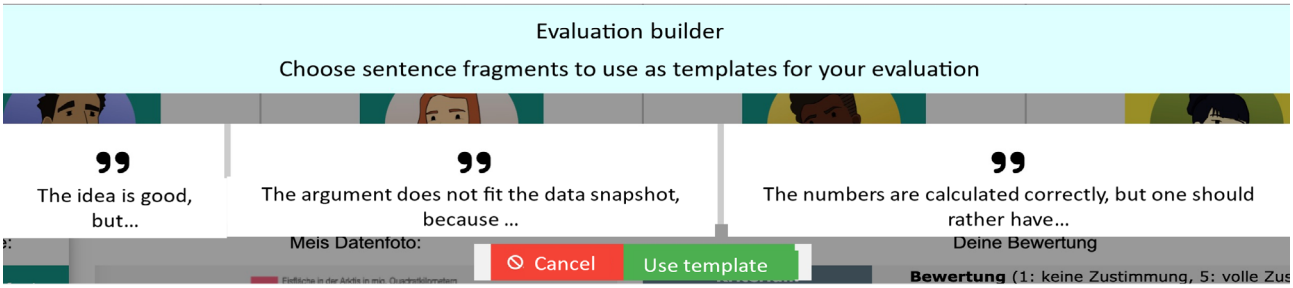


Figure 7: Scaffolds for formulating critique of arguments

Story world

At the very beginning, before starting the work on the learning environment, the interviewer (I) asks the students whether the phenomenon of Arctic sea ice is familiar to them.

I Did you already talk about Arctic sea ice in school? About the Arctic, about Arctic sea ice, or something like that?

C A bit in geography class.

D Yea, in geography.

I And what do you know about it?

C Well, it's a bit critical, what we are doing wrong with things like flying.

I What do you mean?

C That so much ice is melting at the polar caps. And with all those animals.

This excerpt can illustrate the starting point of the design experiment. Both students show some familiarity with the phenomenon of Arctic sea ice decline, but in a very rough way: the melting ice is not specified further, and links between melting ice, flying, and animals are not made explicit.

The story world aims to build on these foundations to establish a more robust context knowledge that can be later used for investigating data and reflecting on argumentation. As their first story, the students read the fictional interview with a polar researcher. In this interview, the researcher provides information about the Arctic seasons, in particular that temperatures vary between the seasons, and that ice melts during the summer. In the following excerpt, the students describe their reaction towards the interview.

D Well, it said that during the summer the ice melts, but it didn't say anything about the ice growing again in winter. And they talked about eternal ice, but…

C But it doesn't always stay the same […] well, but minus 30 degrees in the winter, it could well be possible that ice is created.

I Mhm.

C It's like, far below freezing.

Here, the design principle of building context knowledge first takes effect. Whereas in the very beginning, the students formulated an idea concerning a general "melting" of polar ice, this idea becomes differentiated through the ideas of summer and winter. Dominik shows that this is a new idea that is not easily understood: the interview does not explicitly state that ice regrows in winter. Only after some seconds of thinking, Cedrik comes to the conclusion that a regrowing ice could be explained by the low winter temperatures. This shows that building context knowledge is an active effort by the students, initiated through the design element of the interview story.

Data world

After engaging with additional stories, the students continue to the data world. Before they choose a dataset, the interviewer asks the students how one could investigate the claim that Arctic sea ice is declining. In their answer, the students describe a procedure of taking photos of Arctic sea ice at different times and to compare them. Afterwards, Cedrik chooses the larger dataset with data from 1980 to 2020 for a closer exploration.

I Okay, and why are you choosing this one [*points to the larger 1980 to 2020 dataset*]?

C Because you can make better evaluations with a larger time span. If you only have something like a few months, then you cannot see a big change.

I Why? Why is it that you cannot see it?

C Because for one, one photo could have been taken in winter, and the other in summer.

In this excerpt, it is notable that the students do not justify the use of a larger dataset through theoretical considerations, for example about sampling, but through using their contextual knowledge. Because they know through the story world that the Arctic is influenced by summer and winter, they hold that a larger time span has to be observed, so that no unfair comparisons between summer and winter are made. This idea is not directly

related to the actual dataset at hand. After all, the dataset comprises data from 1980 to 2020, which would include multiple summers and winters. Nevertheless, the students draw on their context knowledge to justify the need for a larger dataset. In this way, the design principle of building context knowledge can influence not only the students' knowledge of the phenomenon, but also their approaches to dealing with data.

Afterwards, the students investigate the data in a table. This is a challenging activity, as the table contains an entry for the minimum ice area for each month for the years 1980, 1990, 2000, 2010, and 2020, and it cannot all be displayed simultaneously.

D Now I see that 2010 it's a lot less ice, but 2020 it balanced out again. There were some numbers more.

I Please explain again. Which numbers do you mean? There are different numbers for 2010 and 2020.

D Well, most of the time, at the last places, there was only like, point six or something. And here then there was a bit more.

C Yeah. From January 2010 to 2020, I think there was 13.5 and then 2020 it was 13.6

[…]

C But when you look at May, it sank a bit. It's different.

D Yes, also in August, 5.5 [*points to 2010*] and 5 [*points to 2020*].

In this excerpt, the students have difficulties expressing a change in the Arctic sea ice. Dominik actually observes that after a brief decline in 2010, the ice recovered in 2020 (it "balanced out again"), which Cedrik explains through a rise that could statistically be considered negligible (2010 "there was 13.5 and then 2020 it was 13.6"). Still, to the students, this seems to be a relevant difference. At the same time, they observe a decline of Arctic sea ice in specific months like August from 5.5 to 5. These two rivaling observations remain unconnected in the students' reasoning. One possible reason could be the missing contextual considerations. The students are fully submerged in the investigation of data (navigating the table) and models (absolute differences between two specific months). At this point in their investigation, ideas relating to the phenomenon of Arctic sea ice, for example summer and winter months, are not articulated.

Afterwards, the students investigate the line graph.

C Now you see really clearly how it changed.

I What do you see?

C That 2020 it's low, like, in the direction of August or September it's very low.

D Yes.

I And where do you see it? Because in 2020, it's also around 14. That seems quite high.

C […] Well, in the time frame between June and December, there it deviated from 2010.

The line graph seems to allow the students to use models of higher complexity. While investigating the table, the students used models of absolute difference between two specific months (e.g. May 2010 and May 2020), they now use a model that compares whole time frames ("between June and December, there it deviated from 2010"). As before, these observations are still grounded in the language of data and models, and the phenomenon is missing. However, when creating their data snapshot, the context knowledge resurfaces (Figure 8). Here, they use a box as a model to signify the summer months in which the decline of Arctic sea ice becomes apparent—knowledge which they first gained from an info card (Figure 3). Thus, they combine model with knowledge of the phenomenon.

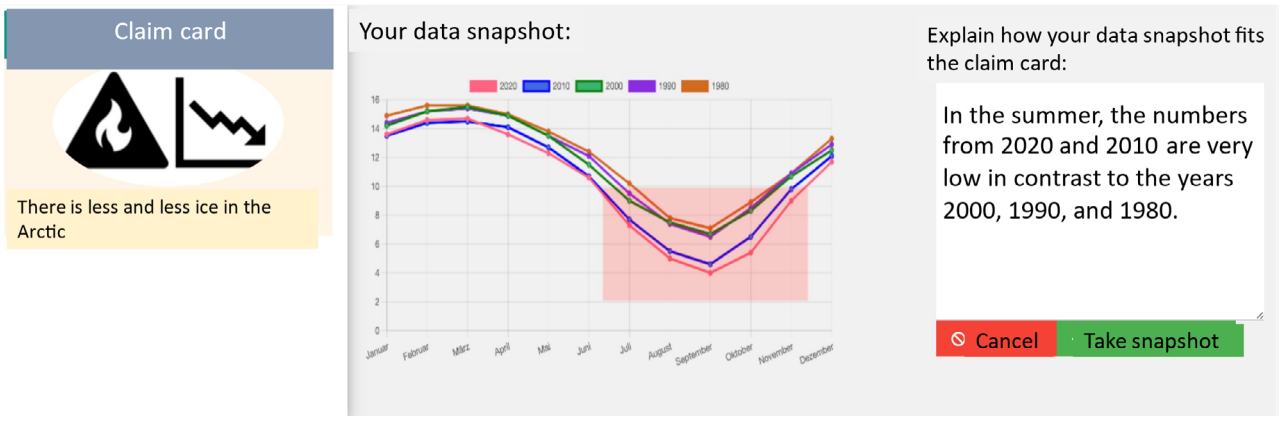


Figure 8: Cedrik and Dominik's data snapshot explicates the links between claim, model, and phenomenon

Argument world

In the argument world, the students engage with Mei's fictitious argument (Figure 6) that uses the data from two years to claim that Arctic sea ice recovers each time, and thus there is no need to worry.C When I look at the diagram, I wouldn't say that it's correct, that argument

I Why?

C Because you can clearly see that, in January, it really is a little bit higher than in May 2020 [...]

In their first reaction, Cedrik clearly rejects Mei's argument. He does so by observing that, although Mei claims that the ice is not decreasing, in her model, you can see that there actually is a decrease from January to May. In this way, Cedrik focuses on the internal coherence between Mei's model and the text of her argument. Therefore, this does not yet constitute a true reflection on her argumentation, as no relation to external aspects of the phenomenon, data, model, or text are provided.

This changes as the pair utilizes the provided scaffolds for reflection:

C Well, 'the data snapshot fits the argument.' I mean, if you look at the short time span, then it fits, but it still is wrong somehow. I don't know how to express it. The argument basically fits the diagram. But if you would have used a different one, with a larger time span, then it definitely wouldn't fit.

The reflection scaffold "the data snapshot fits the argument" prompts a differentiated reflection for Cedrik. Contrary to earlier, he observes that the internal fit between model and text is appropriate ("the argument basically fits the diagram"), but that the argumentation is limited because of its short time frame. His statement that additional data would reveal that the Arctic sea ice is actually declining ("with a larger time span, then it definitely wouldn't fit") can be considered a reflection on Mei's argumentation, as knowledge of external data is related to the argument at hand.

A different reflection scaffold then prompts a different kind of reflection:

C 'Data snapshot and argument are convincing.' No, no, I don't even have to think about this one *[rates the item as 1 – strongly disagree]*.

I *[laughs]* Okay, you mentioned this already.

D Yeah, no *[agreeing]*.

C *[hesitating]* Well, but think again. Basically, it's a scam, but it's not fake. [...] I mean, it feels like a scam. Because if you would give it to a 'relatively uneducated person' *[makes air quotes]*, then they would even believe it.

The differentiation that Cedrik undertakes in this excerpt can be interpreted as Cedrik holding that Mei's argument is grounded in data, and thus "not fake," but still normatively an argument that should not be made ("a scam"). He grounds this consideration in a different type of context knowledge: by drawing on knowledge of how knowledge is distributed in society (there are "relatively uneducated persons") and how argumentation can convince people even of harmful ideas, he draws on external knowledge of social phenomena. Thus, this again presents a reflection of Mei's argumentation, but this time using external knowledge of the phenomenon instead of external knowledge of data.

Summary

In Cedrik and Dominik's progress through the three worlds, the design principles fundamentally influenced the students' learning processes (see overview in Table 1). The design principle of building context knowledge could be considered the most fundamental design principle, as it influenced all three worlds. In the story world, the students could build on their existing context knowledge, which provided a strong motivation. In the data world, context knowledge of Arctic summer and winter helped the students to grasp the complex patterns observed in the sea ice data. And in the argument world, context knowledge of social phenomena provided a lens for reflection on argumentation. In the data world, the students were supported by the tools available to them for creating models as evidence for a claim, which was specified through the claim cards. This provided a goal for the investigation of the data, and the box tool allowed the students to create a model for capturing their individual perspectives on the observed summer and winter patterns. Finally, the scaffolds provided for the design principle of eliciting and supporting reflections on argumentation helped the students use external aspects for reflecting on the argument at hand, thus creating more mature reflections. These reflections on an argument utilizing a quite small dataset were aided by their own experiences with the larger data set: because they observed Arctic sea ice decline in a larger data set, they had knowledge of data and models that were external to Mei's argument relying on just two years of data, which allowed the students to articulate reflections on data and models.

Design principle	Effects on analysis and reflection
Build context knowledge	Motivation due to connection to students' individual knowledge Provide aspects of phenomena for interpreting data Support students' creation of models by providing possible meanings for observed patterns in data Support students' reflections by providing possible aspects of phenomena that are external to an argument
Focus on models as evidence for claims	Strengthen students' arguments by explicitly linking model to text Using tools like boxes to highlight individually perceived relevant patterns without access to formal models
Elicit and support reflections on argumentation	Change perspectives from considerations on internal relations of an argument towards reflection on external domains. Increased differentiation in the evaluation of argumentation

Table 1: Summary of effects of design principles

Discussion and conclusion

The statistical and mathematical contents of modern media pose high critical demands on readers (Gal & Geiger, 2022). This creates a demand for developing statistical literacy, but research discourse in statistics education has provided few guidelines for designing learning environments that are suited to this task (Büscher, 2022). This chapter has reported on a design research study that identified three theoretically and empirically grounded design principles for developing statistical literacy: build context knowledge, focus on models as evidence for claims, and elicit and scaffold reflection on argumentation. The empirical insights into the learning processes of two students working with a digital learning environment that implements the three design principles show how the students' construction of models and their reflection on argumentation is enhanced by these three design principles.

These observations of the working mechanisms of the design principles can provide some additional insights into the points on models and modeling raised by other researchers. Podworny and Frischemeier (**in this volume on page 15**) show how students' conceptions of data as models revolve around data as information and description. In light of this study, the role of data and models as domains of argumentation which are constructed in a purposeful effort to communicate viewpoints and to convince others seems to be unfamiliar to young students. However, as Gal (**in this volume on page 91**) shows, citizens need to react not only to simple descriptions, but to statistical information that is "already digested" and is provided to support a specific viewpoint or policy decision. The two design principles of focusing on models as evidence for claims and eliciting and supporting reflections on argumentation aim to support students in dealing with this demand of analyzing and reflecting "digested information." One way to support this ability is to pay attention to context knowledge. Bielik (2024, **in this volume on page 33**) sketches an approach in which students use rich technological tools to create models, and highlights that during the process of refining models, it is important to frequently revisit the phenomenon to evaluate the model fit. The design principle of building context knowledge supports this revisiting of the phenomenon by providing the knowledge needed to analyze and reflect model fit regarding the phenomenon. Future studies could build on these identified design principles to investigate learning environments for different grades, topics, and contexts.

References

Ben-Zvi, D., Gravemeijer, K., & Ainley, J. (2018). Design of Statistics Learning Environments. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International Handbook of Research in Statistics Education* (p. 473–502). Springer.

Bielik, T. (2024). **"Supporting students' modeling and data practices by engaging with digital tools" in this volume on page 33.**

Büscher, C., & Prediger, S. (2019). Students' Reflective Concepts when Reflecting on Statistical Measures—A Design Research Study. *Journal Für Mathematik-Didaktik*, *40*(2), 197–225. https://doi.org/10.1007/s13138-019-00142-2

Büscher, C. (2022). Design Principles for Developing Statistical Literacy in Middle Schools. *Statistics Education Research Journal, 21*(1). Article 8.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21. https://doi.org/10.1007/BF00988593

Engel, J. (2024). **"Some reflections on the role of data and models in a changing information ecosystem" in this volume on page 101.**

Engel, J. (2017). Statistical Literacy for Active Citizenship: A Call for Data Science Education. *Statistics Education Research Journal, 16*(2), 44–49. https://iase-web.org/documents/SERJ/SERJ16(1)_Engel.pdf?1498680968

Gal, I. (2024). **"What do citizens need to know about real-world statistical models and the teaching of data modeling" in this volume on page 91.**

Gal, I. (2002). Adults' Statistical Literacy: Meanings, Components, Responsibilities. *International Statistical Review*, *70*(1), 1–25.

Gal, I., & Geiger, V. (2022). Welcome to the era of vague news: A study of the demands of statistical and mathematical products in the COVID-19 pandemic media. *Educational Studies in Mathematics*, *111*(1), 5–28. https://doi.org/10.1007/s10649-022-10151-7

Gal, I., Ridgway, J., Nicholson, J., & Engel, J. (2022). Implementing Civic Statistics: An Agenda for Action. In J. Ridgway (Hrsg.), *Statistics for Empowerment and Social Engagement* (S. 67–96). Springer International Publishing. https://doi.org/10.1007/978-3-031-20748-8_4

Garfield, J. B., & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice*. Springer.

Gravemeijer, K., & Cobb, P. (2006). Design Research from the Learning Design Perspective. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. M. Nieveen (Eds.), *Educational Design Research: The design, development and evaluation of programs, processes and products* (p. 45–85). Routledge.

Jablonka, E., & Bergsten, C. (2021). Numbers don't speak for themselves: Strategies of using numbers in public policy discourse. *Educational Studies in Mathematics, 108*, 579–596. https://doi.org/10.1007/s10649-021-10059-8

Makar, K., & Rubin, A. (2009). A Framework for Thinking about Informal Statistical Inference. *Statistics Education Research Journal*, *8*(1), 82–105.

Podworny, S. & Frischemeier, D. (2024). **"Young learners' perspectives on the concept of data as a model: what are data and what are they used for?" in this volume on page 15.**

Prediger, S., Gravemeijer, K., & Confrey, J. (2015). Design research with a focus on learning processes: An overview on achievements and challenges. *ZDM*, *47*(6), 877–891. https://doi.org/10.1007/s11858-015-0722-3

Skovsmose, O. (1994). *Towards a Philosophy of Critical Mathematics Education*. Springer. https://doi.org/10.1007/978-94-017-3556-8

Skovsmose, O. (1998). Linking mathematics education and democracy: citizenship, mathematical archaeology, mathemacy and deliberative interaction. *ZDM, 30*(6), 195–203.

Stephan, M., Register, J., Reinke, L., Robinson, C., Pugalenthi, P., & Pugalee, D. (2021). People use math as a weapon: Critical mathematics consciousness in the time of COVID-19. *Educational Studies in Mathematics, 108*(3), 513–532. https://doi.org/10.1007/s10649-021-10062-z

van den Akker, J. (1999). Principles and Methods of Development Research. In J. van den Akker, R. M. Branch, K. Gustafson, N. Nieveen, & T. Plomp (Eds.), *Design Approaches and Tools in Education and Training* (p. 1–14). Springer. https://doi.org/10.1007/978-94-011-4255-7

Weiland, T. (2017). Problematizing statistical literacy: An intersection of critical and statistical literacies. *Educational Studies in Mathematics*, *96*(1), 33–47. https://doi.org/10.1007/s10649-017-9764-5

Wild, C. J. (2017). Statistical Literacy as the Earth Moves. *Statistics Education Research Journal, 16*(1), 31–37. https://iase-web.org/documents/SERJ/SERJ16(1)_Wild.pdf?1498680942

Zieffler, A., Garfield, J., Delmas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, 7(2), 40–58.

From representation to transformation: rethinking modeling in computer science education

Carsten Schulte

University of Paderborn
carsten.schulte@uni-paderborn.de

Models simplify the world; they make a complicated reality more tractable. Models can reduce a rich, multifaceted phenomenon to a set of data, an intricate relationship to a function, or murky uncertainty to a sequence of stochastic results we can simulate. In this chapter, we consider the role of modeling in software development, and the impact modeling should have in computer science education. The most straightforward use of models in this field is that we often create software to represent something in the real world, that is, software acts as a model for a real-world event or process. This is analogous to the way a statistics or data-science educator might think about modeling, where models represent the objects of our investigation. In this chapter we consider a representational view of modeling in computer science, but also what I will call a transformative view: that implementing and using a piece of software changes our understanding of the task, what is possible, and what is important. While the former view is closely connected to data and algorithms, this latter view is more closely tied to software design. In this chapter, therefore, we will describe the representational view of modeling in more detail, and see why including the transformative view is so important. This leads us, in turn, to intriguing implications for education in software development and design.

The term "modeling" can be interpreted in numerous ways. While this variety is reflected in the broader computer science education discourse (e.g., Caspersen, 2022; Grgurina, 2021), in practice there is one prevalent understanding of the term, which is strongly associated with problem solving, algorithmic thinking, and computational thinking.[1] In the following, I first discuss the typical notion of modeling that is predominant in computer science education, where modeling is closely associated with the development process of software programs and algorithms. I then complement this understanding with a wider perspective on system design and conclude that, today, the concept of modeling should be extended towards a wider view related to design.

As a starting point for discussion, I briefly reflect on how modeling is typically understood in computer science education. While there are several branches of computer science dealing with models and modeling with quite a variety of connotations, computer science education mostly focuses on modeling in the context of programming and software development. As developing software is a highly complex and interwoven task, software projects involve many people who cooperate with the goal of producing a working piece of software. Many important educational aspects are involved: the ability to plan, to be persistent in constructing something, to learn to cooperate with others, and many more so-called "soft skills." In this context, the term model is often used as the visual representation of a plan and hence is tightly connected to aspects of the process of software development. In the first phases of software development, the focus is on understanding the problem and analyzing it. Therefore, the initial modeling stage is creating a model of the problem. In later stages, this model becomes a blueprint for the software to be built. The latter implies that the model must be adapted, because additional technical details of the implementation must be considered

1 One can also see problem-solving in the context of computational thinking as a form of modeling. For instance, Kallia et al. (2021) relate computational thinking in computer science education to modeling in math education.

(beyond the initial representation of the problem), which include properties of the computer system on which the software is to be executed. Therefore, one can differentiate between the two products of the initial and subsequent modeling stages in the following way: The initial product can be seen as a model of the problem, and the later product as a model for the software.

The goal of this paper is to argue that there are different implicit connotations behind these *models of* the problem and *models for* the software, which should be made more explicit. The first one I will label as the *representational* view of models and modeling—as opposed to a view that I will call *transformative*.

The representational view relates to the first phase of software development: the analysis of the problem. An explicit focus on this view became quite popular in computer science education research in the work of Hubwieser and many others (e.g. Hubwieser, Broy & Brauer, 1997; Breier and Hubwieser 2002). It aligned with the typical view in the sciences, where models are mostly construed as representations of the world. While such an interpretation is indeed an interesting and important perspective, the representational view has its limitations: in computer science education, the contexts of use, the societal contexts, and the ethical questions which arise when using the software should be taken into account, as well.

In the representational sense, a model is an image of the world, and software is an implementation of the model. However, when the software is ultimately used, it has an influence on the world. For example, word processors generally allow additional digital features not possible before. Prior to the implementation of such software, text had to be either written down using analog means by hand or by using a typewriter. The result was a fixed product, and altering its format necessitated re-creating the entire text. Digital word processors, on the other hand, allow the formatting of text to be manipulated after it has been written. When templates or styles are used, the digital potential goes even further, as they define the look of certain types of text, independent of where they are in the document. One could, for example, change the look of all headings at the same time. In analog media, changing the look of all headings would be a very tedious and time-consuming operation including cutting the text into pieces and partially rewriting it. Such features cannot be designed based on a model that merely represented the reality the preceded the software development (i.e., based on a representational model of a typewriter, see Winkelnkemper and Schulte (2023) for an extended discussion).

As a result, the world which the model was supposed to represent changes due to the use of that very product that the model served as the basis of. This means that the model immediately became invalid as soon as it became effective. To deal with this discrepancy—and that is the argument of this paper—modeling should not only be seen as the creation of a static or fixed representation (a model *of*), but also as something which has a transformative character (a model *for*). In the next section I will discuss the representational view in more detail, before supplementing it with a transformative view on modeling.

The representational perspective on modeling in computer science education

In computer science education, modeling is classically seen as an integral phase of the process of constructing and developing software, as it encompasses a systematic initial analysis of a task or problem. Software development in turn has played a major role in the debate on establishing computer science as a proper school subject in many countries (e.g., Hubwieser et al., 2015). In this context, the focus in computer science education has shifted from mere coding towards developing thinking skills, specifically for problem-solving through algorithmic thinking (e.g., Gal-Ezer et al., 1995). Nowadays, these thinking skills are often referred to as computational thinking. The focus has turned to those phases of the software development process where problem-solving and related thinking skills are prominently featured.

Breier and Hubwieser (2002) and Breier (2005) argue that computer science is concerned less with computers (in the sense of engineering) than with information, and hence should rather be called "informatics" or "information science" to better express the central role of the subject in education. In my interpretation, this alignment with the sciences is supposed to stress that informatics, and hence informatics education is not concerned with technology, but rather with understanding the world. While in other sciences, experiments are the core method of gaining new scientific insights, in informatics modeling is at the core. Informatics therefore becomes the science of producing models which represent the important features of the real world. In consequence, the scientific progress that characterizes computer science is the constant development of new, often graphical representations of models (e.g., Hubwieser, Broy, & Brauer, 1997).

According to Breier and Hubwieser (2002), the central activity and learning goal for students, hence is to "evaluate software as an informatics model developed following analysis of a problem as a reduced image of reality." (p. 36). Following these notions, the act of modeling is a process that feeds and thereby enhances

thinking abilities, while also representing the nature of the discipline, and hence implicitly, of the school subject. In this view, it is important for students to understand that computer science is not about computers, but rather about viewing the world through the lens of information, and that this lens produces representations of reality on a par with models developed by other sciences.

The perception of computer science as informatics therefor similarly reflects the representational view of modeling, mentioned before when discussing its link to software (and algorithm) development. Here too, the focus is usually put on the initial steps, where a model is developed as a representation of the problem, and how the model is then implemented is deemed far less interesting. Hubwieser even explicitly warns against focusing on the implementational stages. While coding may increase the motivation of learners, according to him, it is of no inherent value and hence should only be allowed as an implementation of a previously developed model (Hubwieser et al., 1997, p. 116ff).

Overall, this perspective and conceptualization of modeling in computer science education forms a coherent world view: The focus is on analyzing a given real-world problem, where "real world" typically refers to an initially nondigital situation of which core aspects are analyzed and subsequently digitized in the form of data structures and algorithms in a graphical notation, such as the Unified Modeling Language (UML), by determining and depicting its structure in a class diagram, or its inner processes as sequence diagrams.[2] The actual implementation only follows as an optional step of filling in technical details to transform the graphical model into actual running source code.

Discussion and criticism of the representational modeling view

At first glance the representational understanding of modeling and the associated role of models and graphical model languages are a sound basis for the school subject of computer science. At a second look, however, they are peculiarly limited in their scope. They strongly focus on the initial phases of idea generation and (problem) analysis, while contributing relatively little to the evaluation and reflection of the outcome.

In the typical science view that models are mainly a tool to understand the real world by creating reduced images of it, models exist mostly to represent reality. In that view, the quality criteria must be whether this image is true or distorted. A good image therefore tends to be construed as objective, neutral and value free. However, the software that is based on such an assumed neutral and objective process has the potential to change the world we live in and have an impact on almost anything and anybody (Magenheim & Schulte, 2006; Rahwan et al., 2019).

For example, Magenheim and Schulte (2006) argue that software development should be understood as developing socio-technical information systems (SIS), including the idea that developing a socio-technical system has to include not only the initial modeling, but also predicting and designing future interactions of humans with the technical system; associated changes in interactions between humans and the technology; and interactions between humans and humans in such an SIS. Therefore, the effects of a technical system cannot be described in isolation but only when also taking its social context into account. They present some general implications and effects of socio-technical informatics systems considering the abstract contexts of production, distribution, education, health care, entertainment and leisure, research, military, and e-democracy (Magenheim & Schulte, 2006). While Magenheim and Schulte focus on traditional software development and products, Rahwan et al (2019) highlight that modern systems based on machine learning have an even more profound effect on the social contexts they are used in, but they nevertheless refer to these systems as algorithms. They provide a summary of examples from the literature on the impact and influence in several areas, which include:

> "…influence the information seen by citizens. […] shape the cost of products differentially across consumers […] shape the dispatch and spatial patterns of local policing […] affect time served in the penal system […] shape romantic matches for online dating services. […] increasingly substitute for humans in the raising of our young and the care for our old. […] increasingly likely to affect collective behaviors, from group-wide coordination to sharing. Furthermore […] machines could determine who lives and who dies in armed conflicts…" (Rahwan et al., 2019, p. 478).

Modeling as described above is blind to these impacts.

The limits of the narrow representational view of modeling described above have been discussed in computer science itself. A prominent example of this is the identification and classification of object-oriented software design patterns. The book by Gamma and colleagues (1995) was (and is) probably the most prominent voice of the idea of general architectural patterns. In the introductory chapters to their collection of patterns, the authors state that "strict modeling of the real world leads to a system that reflects today's realities but not

2 The Unified Modeling Language can be seen as de facto standard for object-oriented modeling in the academy and also in K–12 education as it provides a connected family of graphical models to describe different facets of a system.

necessarily tomorrow's. The abstractions that emerge during design are keys to making a design flexible." (Gamma et al., 1995, p. 24).

Since Gamma and colleagues published their catalog of patterns in the 1990s, several trends in software engineering have emerged that highlight and advocate for the continuous, flexible, and adaptive nature of software development. Evidence of these can be seen in popular buzzwords like agile software development, continuous integration, or DevOps. A more theoretical view and underpinning of why ongoing changes are often needed and are successful can be found in Lehman (1980) or Wegner (1997). Their argument basically is that software is embedded in a context of use, so the inputs or interactions of human users with the system lead to unforeseeable changes which create a need to align the old, original model with new affordances based on these interactions. A change in the context of a software system hence often induces the need to adapt that system.

In addition to these arguments, one can also argue that, as a software system is designed to have some effect (making something more efficient, more precise, or allowing for something genuinely new), the use of the system influences how things are done. In other words, the model that has been implemented was based on a world that no longer exists when the system is put in use. Software which is used is, so to speak, always outdated and hence probably needs to be updated continuously. In practical terms, this means that ideas which seemed useful may not be so useful in practice, and that experiences gained when using the system lead to new ideas. In consequence, the system is likely to be changed—but then again, when the changed system is being introduced, it again creates a world that is different, triggering the onset of another round of new wishes or demands that the software needs to be refined to attend to. This situation of an ongoing need to adapt leads to high interrelations between a software system and the world it is part of—and therefore calls for ideas like agile processes and gives rise to new approaches which see the development and the operation of a system as an integrated endeavor. In this view, it simply does not make sense to try to build a well-defined definite "solution." It is instead more useful to follow a flexible and agile approach to produce small increments in rather short interactions and use the feedback of users for the next development steps (Beck 2001).

A new perspective on modeling as a form of design

Based on the insights in the field of software development and consequently in the field of computer science, I will now suggest a new perspective or a widened picture on modeling and its role in computer science education. I will start with a discussion of ideas from the discipline of science, and afterwards offer implications for computer science education in schools. For this scientific debate, I draw almost exclusively on the work of the theoretical computer scientist Bernd Mahr. From a computer science perspective, Mahr has extensively explored the concepts of models and modeling (e.g., Mahr, 2009; 2011[3]), along with their respective functions. One of Mahr's key insights is that being a model is not an inherent property of some artifact but a form of purposeful ascription and use. Practically any form of representation, object, or even a person in this sense can serve as a model. Mahr introduces the concept of 'cargo' to provide a better understanding of what it means when something is used as a model: the objective of a model is to transfer information or insight—the cargo—from the original (A) to a destination (B) through the model. This objective may involve making certain aspects explicitly visible in the model that are not immediately evident in the original and, in the process, learning from the original or transferring something from it so that the model and hence the derived product acquires a specific property or quality. One notable property of this understanding of models which makes it distinct from the models we discussed before is that models in Mahr's sense serve as an intermediary step or a means of communication within the development process, rather than being its ultimate goal.

Intriguing for computer science education is the fact that something being a model is a flexible decision. In his work, Mahr discusses common properties of models, sometimes by redefining them. He includes classical properties like, for instance, models as abstracted, simplified, rescaled, or scaled-down representations of reality. In other cases, he goes way beyond that, which means that models may even possess additional properties, thereby expanding rather than reducing the original.[4] This particular idea, that a model adds something to the original, applies to virtually all computer science models related to software development. In computer science education, modeling is often referred to as breaking down or operationalizing the problem into calculable individual steps. However, from Mahr's perspective, it can be said

3 See also Upmeier Zu Belzen, Krüger and Van Driel (2019) for a discussion of the work of Mahr.

4 See also Lehrer, R., & Schauble, L. (2006) who discuss a similar view on modeling from a science education perspective.

that the model acquires additional properties, for example those of an algorithm, which the (possibly nondigital, vague, non-automated) original lacks. When processes are automated, their formalization is not just a digital version of what was there before, but something genuinely new. Hence, in computer science, models are not meant to simply mirror reality but to transform it. In line with Mahr, modeling can be portrayed differently from an algorithmic or computational thinking perspective: its goal is not just representing and solving a given problem objectively but to expand and shape a new world.

Possible future directions– from a representational to a transformative view of modeling

Today, the classic characterization of a model as a neutral representation of aspects of the real world might still be predominant in computer science education. With the focus on AI and machine learning systems, however, this might change, as what is typically targeted with such systems often cannot be "modeled" in the classical sense. Rahwan et al. (2019), for example, argue that there is the need for a new discipline that aims to understand the behavior of such systems (which they call machines), as during their development no one can foresee their impact. This study of machine behavior, they suggest, could study one individual machine, one type of machines, or even complex ecosystems of humans and machines in their intertwinedness. The authors suggest that evolution is a useful concept to describe the dynamic process of new software systems. They claim that while humans are shaping these systems, they are also being shaped by those systems at the same time. Even when not particularly focusing on machine behavior, this concurrency of shaping and being shaped should probably be one—or maybe even the most important—aspect of modeling for computer science education (Schulte & Budde, 2018). This thought leads to a conceptualization of modeling not only as representing a reduced version of the real world, but as a process that aims at *transformation* of the world, the users, and, in turn, technology itself.

In the above, I have claimed that models play an important role in the development of software design processes. Similarly to how traditional algorithmic systems are developed, machine learning (ML) systems also go through software design processes. Comparing these different design processes helps us better grasp and explain the different problem-solving and modeling approaches, as well as the systems and models they produce. Considering and understanding these different modeling approaches makes it easier to think about and reflect upon what modeling in computer science is all about.

Knowing different modeling approaches becomes even more important in a world of AI and ML systems (Tedre, Denning & Toivonen, 2021). In the development of algorithmic systems, solving the problem necessitates a thorough understanding of the problem itself. Rules are discerned to govern how the problem should be solved, and these rules are then implemented as the solution to the problem in the form of an algorithm. In the case of machine learning, this analytical understanding is mostly absent. The rules of processing are not explicitly fed into the system by humans and, consequently, are not explicitly designed by humans either. Instead of humans, the digital system itself gradually develops these rules, akin to a statistical pattern, based on the provided data. This process is often referred to as "learning" in Artificial intelligence (AI) jargon. During this process, traditional algorithms or code play a relatively small role (Sculley et al., 2015).[5]

When explaining ML systems, the typical explanatory approach and models of explanation used in traditional computer science education can no longer be applied. In the traditional representational approach, one would focus on "modeling" the ML system itself, emphasizing the algorithmic basis and the fundamental operating principles of neural networks and similar technologies to understand the workings of AI from the ground up. This approach would, however, fail to grasp many important aspects of the behavior of existing ML systems. While the algorithmic basis of ML systems is an important part, its functionality depends on many other factors. A large language model like ChatGPT, for example, cannot be explained by describing the underlying algorithm alone as this would neglect the role of training and data selection which is happening on a large scale. Innovations like these necessitate the development of new explanatory models and potentially new pedagogical approaches that expand or even largely replace the classical approach, especially in the context of AI and ML. The resulting explanatory models are much closer to the role of models in science.

5 See also Tedre, Denning, and Toivonen (2021) for a comparison of traditional rule-driven programming and data-driven developments of ML systems.

In a recent literature summary, Sentence and Waite (2022) suggest using the SEAME-model to map the different teaching approaches for AI and ML that exist so far. SEAME addresses the following levels:

- SE The level of social and ethical considerations.
- A The applications level, where we might use, modify, or create applications that have some AI or ML component.
- M The models level, where we train the model with data. Models output recommendations and predictions for use in applications.
- E The engines level, including neural networks, generative algorithms, data structures, etc. This is the most hidden level, which we are not aware of when we use an application with an ML component.

SEAME overall suggests that for understanding an ML system, different perspectives, including probably different perspectives on models, are needed for a holistic and coherent understanding. Let's take a closer look at the last two levels: The engines level can be seen as focusing on algorithms, whereas the models level focuses on data. Considering the discussions above, could one level be characterized by the transformational view, while the other one related to the representational interpretation of models and modeling? Shouldn't the data models be a representative image of the underlying overall situation? For example, in order to make a sensible recommendation for a movie, the associated data model should accurately represent the types of movies available, as well as the taste of the customer. Any transformation induced by such a model would seem to be problematic. This may be even more troubling when we consider ML systems used in medicine or related to matters of employment, for example, in the hiring process.

These thoughts show that there still is a need for the representational view on models and modeling. But even in these cases, at a second glance, such models could and should also be scrutinized under the perspective of the transformative view. The reasons for this are outlined by Matzner (2016), who, with a focus on data, distinguishes between representational and performative data. To give a simple example: When systems to predict job success are being trained on data from the past, that data often show that men are generally more successful than women.[6] The reason for this result, however, is not a reflection of actual capabilities, but of the data itself. To be both useful and fair, in this case the data needs to be transformed (a developer would probably say: cleaned) so that the under-representation of female success stories is remediated, and the system does not reject applications merely based on gender. This transformation, which might be framed as cleaning, removing outliers, or un-biasing data, clearly does not align with the idea of an objective (fixed) representation. It is instead a purposefully-made attempt to remedy something that is considered unjustified and unjust, to transform an aspect of the use context, or maybe even society in general, that deserves to be made explicit and questioned. The transformation puts the focus on the issue itself: How and why does the misrepresentation occur—and maybe even more importantly, why does it sometimes go unnoticed? In these latter cases, one should definitely not be satisfied when the answer is simply that any model must be a true representation of reality, and hence any intervention is not feasible, or ethically valid.

In summary, the discussion on representational versus transformative views on the modeling process in the context of software development, regardless whether the system is based on traditional algorithms or on machine learning, is intended to point out how each can serve different, complementary educational purposes: the representational view of modeling can be educationally useful to focus on the role of data and models and their inherent link to the "real world" as being a representation—a model of the world. Important qualities like fairness, trustworthiness, or the lack of bias can be discussed from this point of view. The transformative approach can then supplement this view as it can help to focus on the usefulness of the model, on its chosen goals and hence on the possible emancipatory potential (or on associated opportunities and threats), and, in consequence, how to cope with them.

6 See O'Neil, C. (2017) for a discussion of such examples (chapter: Ineligible to serve. Getting a job)

References

Beck, K. (2001). *Manifesto for Agile Software Development*. https://agilemanifesto.org/

Breier, Nobert. 2005. "Informatik Im Fächerkanon Allgemein Bildender Schulen–Überlegungen Zu Einem Informationsorientierten Didaktischen Ansatz." *Unterrichtskonzepte Für Informatische Bildung*.

Breier, N., & Hubwieser, P. (2002). An information-oriented approach to informatical education. *Informatics in Education*, 1(1), 31–42. https://doi.org/10.15388/infedu.2002.03

Caspersen, M. E. (2022). Informatics as a Fundamental Discipline in General Education: The Danish Perspective. In H. Werthner, E. Prem, E. A. Lee, & C. Ghezzi (Eds.), *Perspectives on Digital Humanism* (pp. 191–200). Springer International Publishing. https://doi.org/10.1007/978-3-030-86144-5_26

Gal-Ezer, J., Beeri, C., Harel, D., & Yehudai, A. (1995). A High School Program in Computer Science. *Computer*, 28(10), 73–80. https://doi.org/10.1109/2.467599

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (Eds.). (1995). *Design patterns: Elements of reusable object-oriented software*. Addison-Wesley.

Grgurina, N. (2021). *Getting the Picture: Modeling and Simulation in Secondary Computer Science Education*. University of Groningen. https://doi.org/10.33612/diss.190344009

Hubwieser, P., Broy, M., & Brauer, W. (1997). A new approach to teaching information technologies: Shifting emphasis from technology to information. In D. Passey & B. Samways (Eds.), *Information Technology: Supporting change through teacher education* (pp. 115–121). Springer US. https://doi.org/10.1007/978-0-387-35081-3_14

Hubwieser, P., Giannakos, M. N., Berges, M., Brinda, T., Diethelm, I., Magenheim, J., Pal, Y., Jackova, J., & Jasute, E. (2015). A Global Snapshot of Computer Science Education in K–12 Schools. *Proceedings of the 2015 ITiCSE on Working Group Reports*, 65–83. https://doi.org/10.1145/2858796.2858799

Kallia, M., van Borkulo, S., P., Drijvers, P. Barendsen, E. & Tolboom, J. (2021). Characterising Computational Thinking in Mathematics Education: A Literature-Informed Delphi Study. *Research in Mathematics Education* 23(2):159–87. https://doi.org/10/gnqv9m.

Lehman, M. M.: Programs, Life Cycles, and Laws of Software Evolution. *Proceedings of the IEEE*, Vol. 68, No.9, September 1980.

Lehrer, R., & Schauble, L. (2006). *Cultivating model-based reasoning in science education* (pp. 371–388). na.

Magenheim, J., & Schulte, C. (2006). Social, ethical and technical issues in informatics—An integrated approach. *Education and Information Technologies*, 11(3), 319–339. https://doi.org/10.1007/s10639-006-9012-6

Mahr, B. (2009). Die Informatik und die Logik der Modelle. *Informatik-Spektrum*, 32(3), 228–249. https://doi.org/10/fn4jpp

Mahr, B. (2011). On the epistemology of models. *Rethinking epistemology*, 1, 301–352.

Matzner, T. (2016). Beyond data as representation: The performativity of Big Data in surveillance. *Surveillance & Society*, 14(2), 197–210. https://doi.org/10.24908/ss.v14i2.5831

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. B/D/W/Y Broadway Books.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'Sandy', … Wellman, M. (2019). Machine behaviour. *Nature*, 568(7753), 477–486. https://doi.org/10/gfzvhx

Schulte, C. (2001) Vom Modellieren Zum Gestalten—Objektorientierung als Impuls für einen neuen Informatikunterricht. *Informatica Didactica* 3. http://ddi.cs.uni-potsdam.de/InformaticaDidactica/Schulte2001.htm

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. *Advances in Neural Information Processing Systems*, 28. https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

Sentance, S., & Waite, J. (2022). Perspectives on AI and data science education. in *AI, Data Science, and Young People. Understanding Computing Education (Vol 3)*. Proceedings of the Raspberry Pi Foundation Research Seminars. https://rpf.io/seminar-proceedings-vol-3-sentance-waite

Wegner, P. (1997). Why interaction is more powerful than algorithms. *Communications of the ACM*, 40(5), 80–91. https://doi.org/10.1145/253769.253801

Winkelnkemper, F., & Schulte, C. (2023). Reconstructing the Digital—An Architectural Perspective for Non-Engineers (Discussion Paper). In *23rd Koli Calling International Conference on Computing Education Research*. Association for Computing Machinery.

Tedre, M., Denning, P., & Toivonen, T. (2021). CT 2.0. In *21st Koli Calling International Conference on Computing Education Research*. Association for Computing Machinery.

Upmeier Zu Belzen, A., Krüger, D., & Van Driel, J. (Eds.). (2019). *Towards a Competence-Based View on Models and Modeling in Science Education* (Vol. 12). Springer International Publishing. https://doi.org/10.1007/978-3-030-30255-9

Reimagining data education: Bridging between classical statistics and data science

Ronit Gafny and Dani Ben-Zvi

The University of Haifa, Israel
ronit.gafny@edtech.haifa.ac.il, dbenzvi@univ.haifa.ac.il

The extensive use of big data is fundamentally transforming various sectors, governance structures, and societal norms, leading to significant shifts in our perception and engagement with the world. This chapter concentrates on exploring data models and modeling as educational activities in the intersection of classical statistical education and the domain of data science education. The chapter centers on data models and their diverse applications, spanning from description and explanation to prediction, and encompasses different types of data, both traditional and non-traditional. The primary objective of the chapter is to introduce a Hypothetical Integrated Data Modeling Learning Trajectory (HIDMLT), which constitutes an initial phase within a broader design research initiative. This initiative aims at studying students' reasoning with data models and modeling in both classical statistics and data science. The HIDMLT is based on the application of the Variability, Data, and Phenomenon (VDP) framework (Gafny & Ben-Zvi, 2023).

This chapter begins by outlining the theoretical underpinnings that influenced the creation of the HIDMLT, highlighting distinctions between data science and classical statistics. The key elements, Informal Statistical Models (ISMs) and the VDP framework are introduced. The chapter then presents the complete HIDMLT. The chapter concludes by sharing preliminary insights gleaned from initial implementations of the HIDMLT, shedding light on its potential to enhance pedagogical aspects related to data models and modeling.

Introduction

In recent years, a large part of the world's population has lived in a digital environment embedded in countless applications based on data science and its subfield of machine learning. These applications shape the way we perceive the world around us and even ourselves. There are many ethical, legal, educational, and social questions raised by the impact of data science on our daily lives (Zuboff & Schwandt, 2019). Due to this proliferation of data and computing infrastructure, data science has only recently matured as a research area (Desai et al., 2022). As a relatively young field, it has raised several challenging educational questions, such as: What are the "big ideas" of data science? How do people (experts vs. novices) reason with data science? What are the epistemological foundations on which it relies? What do citizens need to know about how data science works? Such questions can help us identify the skills, knowledge, values, and attitudes necessary for citizens and professionals to develop their ability to use data science products responsibly and effectively.

In the realm of data analysis, both statistics and data science share a fundamental bond as they revolve around drawing inferences and prediction based on data. This intrinsic connection between the two disciplines underscores their close relationship and complementary nature. As a result of this close relationship, educational research in these disciplines should also be closely related. A solid foundation for data science education can be found in building on the already established field of statistics education. Existing results from statistical education research have the potential to inspire innovative methods for teaching data science and identify areas where further research and development are needed to improve data science education. As part of our research, we aim to apply the knowledge and understanding gained in statistics education on informal statistical reasoning to big data and data science settings. In this chapter, we offer a Hypothetical Integrated Data Modeling Learning Trajectory (HIDMLT) which relies on the implementation

of a theoretical framework—the Variability, Data, and Phenomenon (VDP) framework (Gafny & Ben-Zvi, 2023). The VDP framework provides a useful lens to consider various aspects of uncertainty that arise when working with nontraditional data (Noll et al., 2023).

This chapter begins with the theoretical background that inspired the development of the HIDMLT, focusing on the differences between data science and classical statistics regarding data, data models, data modeling, and uncertainty related to both disciplines. An overview of informal statistical models (ISMs) and the VDP framework are provided as key building blocks of the HIDMLT. The second part of the chapter presents the full HIDMLT, including background, a pilot study, the data employed in the pilot study, and a thorough depiction of the suggested learning trajectory. The concluding part focuses on initial findings regarding the HIDMLT and discusses them.

Theoretical Background

The first part of this theoretical section focuses on how classical statistics relates to data science in the educational context. We explain the relationship between the two disciplines and provide an overview of the types of data that each discipline employs. We include a definition of data models and explain the different modeling cultures of the two disciplines and the importance of uncertainty in both of them. We follow with a discussion of informal statistical models (ISMs) (Dvir & Ben-Zvi, 2023) and the VDP framework (Gafny & Ben-Zvi, 2023).

The relationship between classical statistics and data science

Statistics and data science are closely related as they both involve using data to gain knowledge or wisdom (Rowly, 2007). In general, data science emerged from the field of statistics and computer science and gradually developed into a field of its own. Inferences based on statistics, encompassing mathematics, empirical science, and philosophy, have been studied since 1763. A significant increase in computational power and accessibility was witnessed during the "computer age" in statistics, which started in the 1950s, leading to increased use of predictive algorithms. With the advent of big data and technological advancements, data scientists can now collect and analyze large amounts of data. This allows them to extract valuable insights, further cementing data science's role as a separate field from statistics (Efron & Hastie, 2016).

Statistics is defined by the American Statistical Association as "the science of learning from data and of measuring, controlling, and communicating uncertainty" (Wild, Utts, & Horton, 2018). In contrast to statistics, data science as a relatively newly developed field, has various definitions (Lee et al., 2022). All the proposed definitions include statistics as an essential element. Desai (2022) describes data science as the study of information systems (natural or artificial) using probabilistic reasoning (e.g., inference and prediction) implemented through computational tools (e.g., databases and algorithms). In other words, data science is an interdisciplinary field combining statistical methods, computer science, and domain expertise to develop insights from data. Today's emerging consensus is that data science includes statistics as a subset (Donoho, 2017), and statistics is at the core of data science and provides the foundation for many of its methods.

Since statistics is an integral part and is even considered to be the basis of data science, it is possible to develop a world of research pertaining to data science education by building upon the already well-established world of statistics education (Ben-Zvi, Makar, & Garfield, 2018). To do that, we must be aware of the differences between the disciplines. We begin exploring the differences by considering traditional data and big data, data models and modeling, and the role of uncertainty in both disciplines.

Traditional data versus big data

Classical statistical analysis assumes that data is collected using a well-designed sampling scheme, relying on reliable measurements of high quality to provide evidence for a well-defined research problem. In classical probability theory and hypothesis testing, Euclidean data—data that can be easily represented as coordinates—is often applied (Zhang, Liu, & Xiong, 2022). In contrast, modern real-world data is derived from various sources, such as natural language processing, translation, speech recognition, mathematical formulas, computer programs, social networks, transportation networks, sensor networks and automation, as well as biomedical and biomolecular measurements. Many times, this data may not adhere to the Euclidean data models typically used in classical statistics (Zhang, Liu, & Xiong, 2022). Furthermore, in many cases, data are opportunistic data, collected incidentally or as a byproduct of some other activity, rather than being collected intentionally for a specific purpose following a well-designed sampling plan, for example, data collected through sensors, mobile devices, or social media platforms. Even though these data may not have

been collected with a specific purpose, they can still contain valuable information. For instance, data generated from social networks can be unreliable, unsafe, and even false (Holmes, 2017), but can provide valuable insights to medicine, psychology, education, etc.

In the 1990s, technology companies specializing in data analysis introduced the term "big data." Doug Leney (2001) wrote a report for Mata (now Gartner, https://www.gartner.com/) that introduced the concept of big data. This report laid the foundation for defining "big data" without using the term. The report outlined three main dimensions of data management: *volume*, *velocity*, and *variety*. Generally, *volume* refers to the amount of data accumulated, *velocity* refers to the speed at which data is generated and collected, and *variety* refers to the diversity of data types. The evolution of big data has led to the development of new and more advanced data science techniques, as classical statistics had difficulties handling them (Breiman, 2001).

Data models and modeling

In general, models can be viewed as analogies and representational systems, offering a simplified description of a real-world situation (Hesse, 1962). In a sense, models simplify the complex systems or phenomena they seek to represent. Data models differ from other models in that they involve data. Data models can be viewed as methods for organizing, analyzing, manipulating, and explicitly representing data to capture phenomena and specific aspects of the world (Leonelli, 2019), for example, stochastic models. It is important to note that data itself can also act as a model (see Podworny & Frischemeier, **in this volume on page 16**), especially during exploratory data analysis, when researchers and analysts often start by directly examining the raw data to gain insights and identify patterns, trends, and relationships. Modeling is a comprehensive concept that covers the entirety of the process involving the conception, design, and creation of models allowing us to bridge the gap between raw data and the underlying phenomena they seek to represent. Data models can serve various purposes, such as forecasting and explanations (Leonelli, 2019). In a similar vein, the work of Dvir and Ben-Zvi (2018) shows that statistical models might be utilized for descriptive, explanatory, or predictive purposes.

According to the landmark work of Breiman (2001), there are two distinct cultures when it comes to using models to reach conclusions from data. One assumes that a given stochastic data model generates the data. The other uses algorithmic models and treats the data mechanism as unknown. The latter algorithmic approach strongly influences today's data science.

The implications of these cultural differences between classical statistics and data science can be seen in various aspects. The classical modeling approach begins with choosing a simple model (such as the Gaussian model) based on intuition about the mechanism by which the data is generated, and a strong emphasis is placed on the model's interpretability and validity. These models are usually developed to enable a particular type of analysis, namely, generalization from small samples to large populations under rigid data collection protocols (Gould, 2024 **in this volume on page 81**).

On the other hand, algorithmic modeling begins with selecting the model with the highest predictive validation accuracy (such as a random forest), with no regard for the model's explainability at all (Koehrsen, 2019). According to Gould (again, **in this volume on page 81**), these predictive algorithmic models consist of collections of many algorithms that each "votes" on a prediction, and the prediction receiving the most votes is declared the winner. Breiman (2001) argues that algorithmic models are far more flexible, scalable, and accurate for complex big data problems.

It can be observed that classical statistics, with its descriptive, explanatory, or predictive capabilities, is less effective in predicting from big data. Conversely, data science tends to excel in prediction but is relatively less focused on explanation. In examining data science education, it is important to consider the implications of the above differences in the purposes and techniques of data models and modeling.

A key distinction between stochastic data models and algorithmic models is the way they approach uncertainty and patterns in the data, which we discuss next. Stochastic data models use probability distributions to account for randomness in the data, while algorithmic models learn patterns directly from data using computational techniques.

Uncertainty in the realms of statistics and data science

In statistics education, one of the key challenges is teaching students how to deal intelligently with uncertainty (Manor, Ben-Zvi, & Aridor, 2013). Despite its importance in statistics, uncertainty is not sufficiently emphasized in school curricula, and students of all ages have difficulty analyzing uncertainty (Moore, 1990). Uncertainty is a key aspect of statistics, typically accounted for using probability and is no less central in the world of data science. Uncertainty is related to "data" and "chance" as treated by statistics and probability, respectively (Moore, 1990). Statistics tend to focus on randomness-related uncertainty, while probability allows measurement of the level of uncertainty that characterizes the phenomenon itself. There are two types of uncertainty when dealing with classical statistics: statistical and contextual. *Statistical uncertainty* makes it difficult to infer a population from a random sample because two opposing ideas must be reconciled. On the one hand, a sample can represent a population, and the other is sampling variability, which means that even samples of similar size can provide different perspectives on the same phenomenon (Manor & Ben-Zvi, 2015). *Contextual uncertainty* results from conflicts between students' contextual knowledge, specifically regarding what they reflect about the investigated phenomenon and what the data tell them (Manor et al., 2013).

The uncertainty inherent in data science and big data is no less fundamental. The large volume and diverse nature of this type of data often lead to inherent uncertainties. These uncertainties can stem from various sources, such as data inconsistencies, category ambiguity, randomness, partiality, omissions, and structural and organizational issues (Hariri, Fredericks, & Bowers, 2019). The way data is modeled and used can also contribute to other developments of uncertainties.

In data science, it is common to distinguish between *aleatory* and *epistemic* uncertainty when discussing models. Aleatory uncertainty pertains to the concept of randomness, meaning the variability in data caused by random factors. On the other hand, epistemic uncertainty arises from a lack of knowledge or understanding of the situation, also known as a decision-maker's ignorance of the situation (Hüllermeier & Waegeman, 2021).

In conclusion, classical statistics and data science are closely related disciplines that share many similarities but also have distinct differences. One major difference is the type of data each field uses. Classical statistics often focus on smaller datasets. Data science, on the other hand, deals with much larger and more complex datasets in a variety of forms (text, numbers, sound, videos, pictures, etc.). Another difference is the type of models each field prefers. Classical statistics typically rely on parametric models, which require assumptions about the underlying distribution of the data. In contrast, data science often employs non-parametric models and complex algorithms, allowing for more flexibility in handling a wider range of data distributions and structures. Overall, while classical statistics and data science share many similarities, the differences in the types of data and models used, as well as the different perspectives on uncertainty, make them distinct disciplines with different applications and use cases.

Informal statistical models

Informal statistical models (ISMs) form the basis for the data exploration activities that are part of this study and provide inspiration for developing the informal predictive big data model-building activity. As defined by Dvir and Ben-Zvi (2023), an ISM is a purposeful representation that accounts for how the observed variability in data was generated. This representation includes (1) an informal deterministic component, which models some sources of the patterns in the variability observed in the data; and (2) an informal stochastic component, which models some sources of the noise—the observed variability—in the data. The process of informal statistical modeling, which encompasses the creation, development, evaluation, and refinement of ISMs, may result in the generation of preliminary or incomplete versions. These versions might consist of purely deterministic models or models that acknowledge observed variability without addressing its sources.

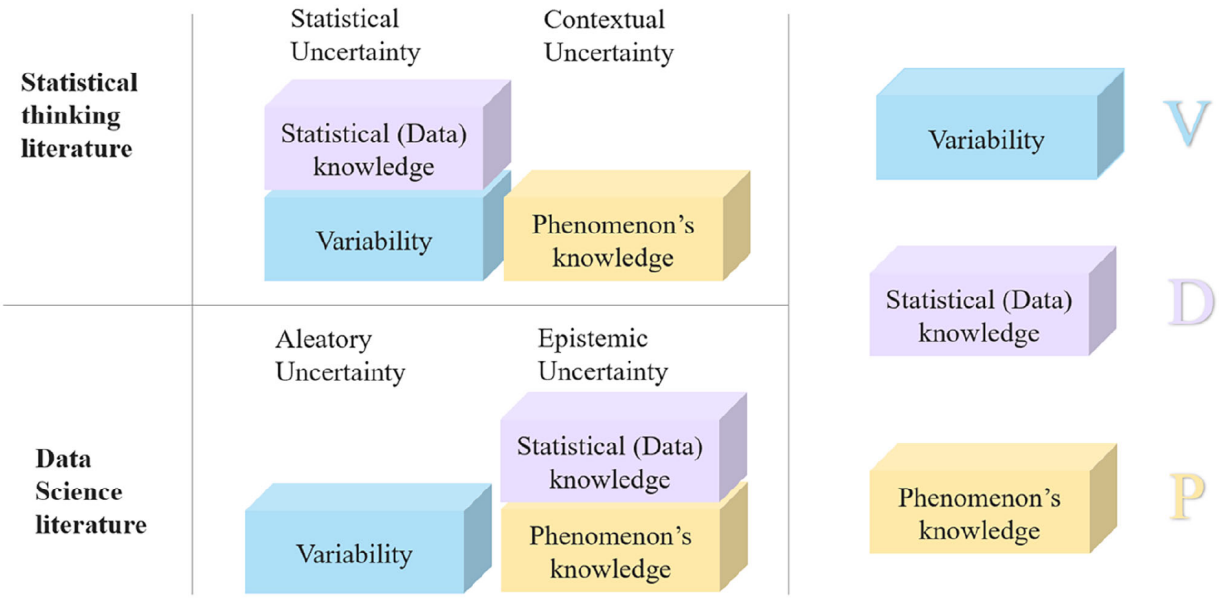


Figure 1: The VDP framework synthesized from different literature distinctions.

The VDP framework

The Variability, Data, and Phenomenon (VDP) framework has been developed to assist in analyzing, describing, and understanding students' articulations of uncertainty during data exploration activities (Gafny & Ben-Zvi, 2023). The VDP is based on two distinct bodies of research: statistical reasoning and data science. As mentioned earlier, two types of uncertainty are associated with data explorations in the traditional statistical reasoning literature: *statistical* uncertainty and *contextual* uncertainty. Statistical uncertainty results from variability in data, while contextual uncertainty results from the phenomenon that is being explored. The conventional classification of variability in data science is different. In predictive models, data science distinguishes between *aleatory* and *epistemic* uncertainty. Knowledge is at the heart of this distinction. Aleatory uncertainty arises from natural, unpredictable, and irreducible variability, whereas epistemic uncertainty arises due to a lack of knowledge concerning the studied phenomenon and the data behavior.

The VDP framework takes these four types of uncertainty and classifies them according to the sources of the uncertainty, as shown in Figure 1.

The result is a three-category model:

1. *Variability* is a common source of statistical uncertainty and aleatory uncertainty;
2. (lack of) knowledge regarding the *Data* is a common source of epistemic uncertainty and might be considered (in a broader sense) part of statistical uncertainty; and
3. (lack of) knowledge regarding the *Phenomenon* is a common source of contextual uncertainty and epistemic uncertainty.

As illustrated in Figure 1, the building block of aleatory uncertainty is variability (Gafny & Ben-Zvi, 2023). The building blocks of epistemic uncertainty are knowledge of the data and the phenomenon. The building block of contextual uncertainty is knowledge regarding the phenomenon. One building block of statistical uncertainty is variability, and we can also consider knowledge regarding the data.

The VDP framework can be a useful tool for analyzing, understanding, and promoting students' reasoning with uncertainty in big data. It proposes a new classification of types and sub-types of uncertainty. We utilize the VDP framework as a bridge that enables the exploration of classical statistics and data science practices. This is due to the central role of uncertainty in constructing data models in classical statistics and in data science. Utilizing this framework may enhance the understanding of beginners regarding the different types of variability that data models encompass. Our proposal is to apply the uniform VDP framework to help students identify uncertainties that may result during data exploration and modeling. Students can enhance their comprehension of the data model by contemplating the elements of uncertainty present in the exploration process and in its resulting data models.

The Hypothetical Integrated Data Modeling Learning Trajectory (HIDMLT)

Background

Deepening students' understanding of variability, data and phenomenon can support their reasoning with the different uncertainties they encounter in the big data context (Gafny & Ben-Zvi, 2023). This study also demonstrates the pedagogical potential of integrating traditional data and big data investigations into a single sequence. The initial HIDMLT was developed as part of a follow-up study to delve deeper into understanding students' reasoning processes during activities that incorporate big data in conjunction with traditional data analysis. The primary focus is on examining how students approach modeling activities in both disciplines.

The current pilot study

The proposed HIDMLT is the basis for pilot research involving a pair of 16 years old female students from Haifa, Israel. The pilot was conducted as part of the Connections (https://connections.edtech.haifa.ac.il) longitudinal design and research project (began in 2005) aiming at promoting young learners' statistical reasoning in a technology-enhanced and inquiry-based learning environment (Ben-Zvi, Gravemeijer, & Ainley, 2018). The pilot case study was based on an extended learning sequence developed for the Citizen Science project "Sleep: One-Third of Our Life" (the Sleep Project), part of the Taking Citizen Science to School (TCSS, https://www.tcss.center) research center. The project deals with teenagers' sleep habits in Israel.

Upon concluding and analyzing the pilot phase, the suggested HIDMLT will be further employed in an upcoming larger-scale study centered on data models and modeling within the realms of classical statistics and data science. This forthcoming study will target a group of around 20 tenth grade data science students.

The data that was used implementing the HIDMLT in the pilot study had been collected as part of the citizen science Sleep Project. About a thousand students were asked to maintain a 14-day sleep diary that encompassed a selection of 20 attributes specifically chosen by the project's researchers. Additionally, each class had the flexibility to augment the data collection by incorporating supplementary attributes alongside the ones initially selected by the researchers.

The pair used two kinds of data sets. The traditional "small" dataset comprised data they collected over a 14-day period through their sleep diaries, encompassing 30 attributes. The "big data" dataset involved 63 attributes. While 20 attributes were uniformly recorded by all students, the remaining attributes exhibited variable completion rates among participants, since each class had chosen its preferred additional attributes to explore. This comprehensive dataset resulted from consolidating contributions from multiple classes engaged in the project, culminating in a dataset containing 16,000 cases (each of them represent one night), collectively representing the contributions of over 1,000 students and 38 classes.

The design methodology

The research methodology used in this study is design research. Design research involves several stages, including preliminary research, prototyping, and assessment. The preliminary research phase begins with analyzing needs and context, a literature review, and developing a theoretical framework. In the prototyping phase, an iterative design process consists of macrocycles of research aimed at refining the intervention through formative evaluation. The assessment phase involves summative evaluation to determine whether the solution or intervention meets the predetermined specifications. This phase often results in recommendations for intervention improvement (Plomp, 2013).

This chapter focuses only on the preliminary design phase and includes a preliminary hypothetical learning trajectory design that was tested on a pair of students during May–June 2023. Simon (1995) introduced the concept of a hypothetical learning trajectory (HLT), which outlines essential components for designing lessons. This trajectory encompasses the desired learning outcomes for students, the activities aimed at facilitating their learning, and conjectures about how the students will learn (Simon & Tzur, 2014). In other words, HLT consists of learning goals, a set of learning tasks, and hypothesized learning process (Apriyanti, Suweken, & Suparta, 2019).

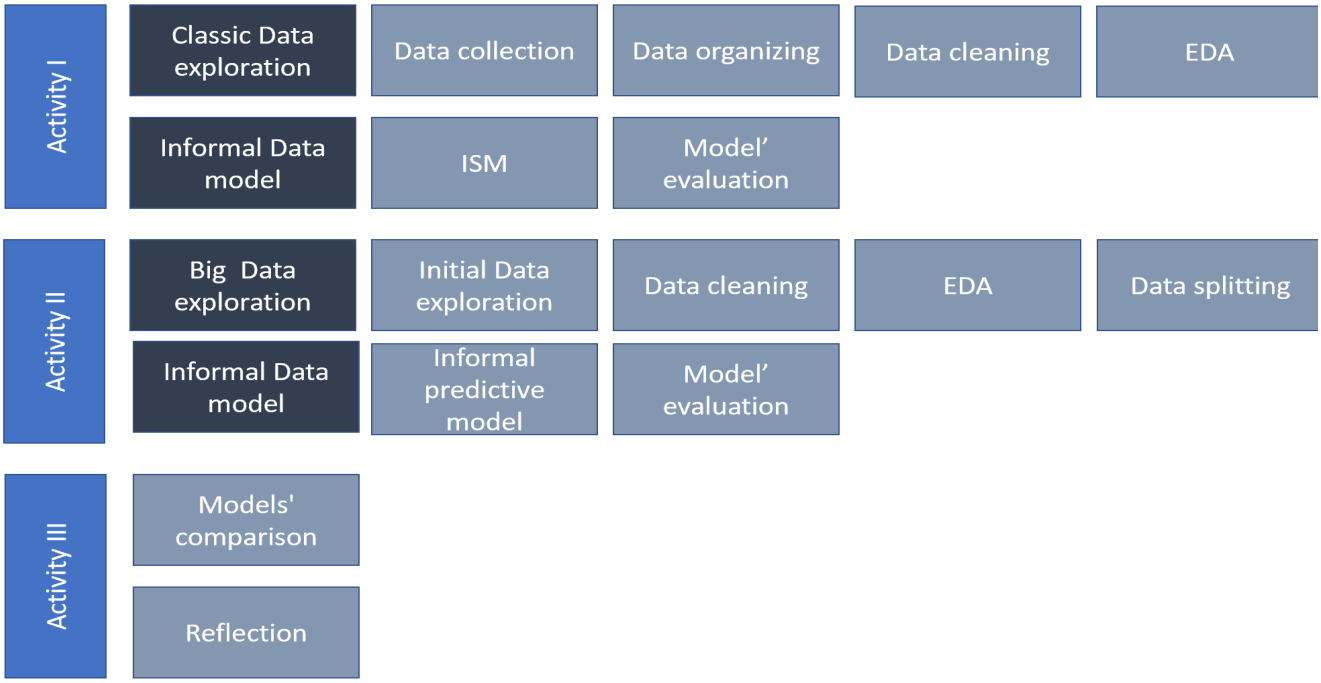


Figure 2: High-level description of the proposed HIDMLT.

The HIDMLT structure

Figure 2 is a high-level description of the proposed HIDMLT. The HIDMLT consists of three parts: a classic data activity, a simulation of big data activity, and a comparison of models. All three activities involve the VDP framework considerations.

The classic data activity. The main goal of this activity is to build a classic data ISM. To achieve this goal, the students are required first to understand the context, perform self-gathering of data, organize the data, and perform Exploratory Data Analysis (EDA, Tukey 1977) with TinkerPlots (Konold & Miller, 2015). During the informal data modeling phase, students are prompted to explore questions that interest them within the dataset. They then proceed to develop a model using the TinkerPlots Sampler tool, which helps them assess the feasibility of extrapolating their findings from the collected data to the broader population. This process involves constructing ISMs (Dvir & Ben Zvi, 2023).

Big data activity. The goal of this activity is to build an informal big data predictive model. To achieve this goal, the students are asked first to clean and prepare the data, perform EDA, and identify patterns and relationships in the data. In this big data activity, during the informal data modeling phase, students are guided to delve into inquiries that interest them in the dataset. Subsequently, they utilize TinkerPlots' sampler tool to create a model. The aim here is to construct a model capable of predicting the behavior of a specific group within the data. To facilitate this, the data is divided into "train" and "test" datasets. Importantly, since students do not employ algorithms or formal predictive models, the models they are crafting are informal in nature.

The model comparison activity. The goal is for students to assess the models' quality. To achieve the goal, the students first assess each model separately and then compare the models they created, their level of certainty in the models, and the different exploration processes that led them to the model.

The detailed HIDMLT

In Table 1, we provide a comprehensive list of activities, sub-activities, pedagogical goals, and students' learning process hypotheses that form the suggested HLT (Simon, 1995). Each sub-task has its own learning objective that emulates an authentic formal practice and serves as a stepping-stone to the subsequent sub-activity. Additionally, the Table includes a recommendation as to how many lessons should be assigned to each activity.

Some of the considerations of uncertainty will arise naturally from the designed assignments. For example, while cleaning the data, considerations regarding the quality of the data and the level of confidence may arise. In other places, or if the considerations do not arise, questions related to the level of confidence during the data exploration and modeling process can help the student to reason with uncertainty. The performance of the activities described above (Table 1) should therefore be scaffolded by a facilitator who is attentive to the students' expressions of uncertainty to guide them through various questions that can expand the resonance of the VDP considerations. Facilitators should be therefore familiar with different types and examples of uncertainty within the VDP framework, including variability uncertainty, data uncertainty and phenomenon uncertainty (Gafny & Ben-Zvi, 2023).

The HIDMLT: Preliminary insights and future directions

The initial pilot case study, although not yet extensively analyzed, has already uncovered some noteworthy findings. Firstly, the active engagement of students in data collection holds considerable importance. Designing the learning trajectory so that students collected their own data that was subsequently integrated into the comprehensive dataset led to a much smoother elicitation of variability, data, and phenomenon (VDP) considerations when dealing with the secondary big dataset, in comparison to the previous study (Gafny & Ben-Zvi, 2023). Additionally, the process of constructing the ISMs posed challenges for the students, suggesting a potential need for a more comprehensive grounding in knowledge related to stochastic data models. Moreover, the shift from exploring "small" data to delving into the realm of "big data" on the same subject seems to ease the transition between these somewhat distinct domains. However, it also introduces specific intricacies and subtleties that merit further examination and deliberation.

The pilot phase serves as preparation for the upcoming study, which will scale up the HIDMLT approach and conduct further research. This larger study aims to enhance the understanding of how students interact with and interpret different data models (including stochastic and real-world models) that serve different purposes, ranging from descriptive and explanatory to predictive, and encompass different types of data (traditional and non-traditional).

The introduction of the HIDMLT serves as a practical example that illustrates the integration of data science and classical statistics in data modeling and related activities. It therefore carries valuable insights for both educators and researchers in terms of the pedagogical integration of various dataset types, the two disciplines' cultures, and diverse data models.

The Minerva Conference 2022 provided a platform for attendees to expand their knowledge of data models. We gained a deeper understanding of what data models are, the various types of models, and the different definitions of the term. Furthermore, the conference allowed for extensive engagement with predictive models in data science, which offered a valuable opportunity to understand the differences between reasoning within data science and classical statistics. This increased understanding and exposure to different perspectives and approaches have opened new avenues for exploring data models and their applications.

Table 1 The HIDMLT includes a list of activities, sub-activities, pedagogical goals, and students' learning process hypotheses.

Stages	Learning Goal and the authentic practice that the activity reflects	Learning Activity	Hypothesized learning process
Classic data exploration. (Two lessons, 90 minutes each)	Data collection—learning about the context of the project and the importance of data quality.	Self-collection of data by students.	Self-data collection will connect students to the subject matter and serve as a basis for creating curiosity. After collecting self-data, the students would be able to compare it with the data of their peers.
	Data organizing—developing an initial dataset. Learn about data structures, databases, and data management tools.	The students gather data from the entire class and consolidate it into a single dataset, which they organize and manage.	The data's organization and consolidation will allow comparison and search for connections. When the data is organized, anomalies and deficiencies will naturally show.
	Data cleaning—learning how to manage and organize data which involves identifying errors, inconsistencies, and outliers and applying appropriate techniques to correct or remove them.	Cleaning the data by identifying and correcting errors, removing missing or duplicate values, and addressing outliers or inconsistencies in the data.	By cleaning the data, students ensure that the analysis is based on reliable and valid data, which can lead to VDP considerations
	EDA—gaining a deeper understanding of the data, identifying potential patterns and relationships, and developing hypotheses about the relationships and patterns in the data.	Creating visual representations of the data, such as histograms, scatterplots, or boxplots, to identify patterns, trends, and relationships. Calculating summary statistics. Examining the data for unusual or unexpected patterns, such as outliers or missing values, and investigating potential sources of these patterns.	Students will be able to identify interesting connections and build a hypothesis. Developing the hypothesis may raise generalization questions that will lead to VDP considerations, especially around sample size.

Stages	Learning Goal and the authentic practice that the activity reflects	Learning Activity	Hypothesized learning process
ISM building (Two lessons, 90 minutes each)	Informal statistical model—developing the ability to develop a statistical model.	Building an informal conjecture model about the population.	Students will raise considerations regarding sample size.
	Model evaluation—cultivating an awareness of thinking and understanding of uncertainty in data models.	Investigating random simulated data samples, growing sample sizes, and inventing methods to compare between samples.	The model evaluation may lead to VDP considerations and questions about how to cope with different types of uncertainty.
Big data exploration (Two lessons, 90 minutes each)	Data—Gaining first-hand exposure to the vastness and intricacy of big data.	Familiarizing with the data.	The students may feel overwhelmed by the amount of data and driven to the next activity to "calm" the awash-in-data feeling (Erickson, 2020).
	Data cleaning—developing an understanding of data quality and how to improve it.	Removing errors, inconsistencies, and duplicates. Handling missing data by imputing values or removing records with missing values. Resolving discrepancies between different data sources.	Clearing and arranging the data may raise VDP considerations regarding data quality.
	EDA—Gaining a deeper understanding of the data, identifying potential patterns and relationships, and developing hypotheses about the relationships and patterns in the data.	Creating visual representations of the data, such as histograms, scatterplots, or boxplots, identifies patterns, trends, and relationships. Calculating summary statistics. Examining the data for unusual or unexpected patterns, such as outliers or missing values, and investigating potential sources of these patterns.	Students will be able to identify interesting connections. The data exploration may raise VDP considerations.
	Splitting data will allow students to evaluate the performance of the future model they will build in the next stage.	Splitting the data into training and testing datasets.	

Stages	Learning Goal and the authentic practice that the activity reflects	Learning Activity	Hypothesized learning process
Predictive big data model building (Two lessons, 90 minutes each)	Pattern recognition—developing an ability to identify associations and patterns.	Deepen the exploration of interesting associations or patterns that would be a base for model building.	Once an association has been identified, the question may arise whether predicting behavior through this relationship is possible.
	Predictive model building—developing the ability to understand how to produce data models and what are their strengths and weaknesses.	Students develop a model for informal forecasting.	Students will raise considerations regarding the model's ability to predict.
	Model evaluation—cultivating an awareness of thinking and understanding of uncertainty in big data models.	Through the application of the model to test data, students will be asked to assess the quality of the model's prediction.	Students will be able to determine whether they can rely on the model and at what level. Considerations related to VDP may arise.
Comparison and reflection on the data models and the modeling process (One lesson of 90 minutes)	Comparing models and reflecting on the modeling process—cultivate an awareness of thinking and understanding of uncertainty in both types of models and deepen the understanding regarding data models and their different objectives.	Comparing the data models produced in the former activities and reflecting on the similarities and differences between the two modeling processes and their objectives.	Students might enhance their understanding of the differences between big data and classic data and deepen their understanding of data models and the uncertainties involved in the exploration and modeling process.

References

Apriyanti, M., Suweken, G., & Suparta, I. N. (2019). The development of hypothetical learning trajectory for linear equation with PISA and scientific approach model consideration. *Jurnal Pengajaran MIPA, 24*(2), 78–86.

Ben-Zvi, D., Gravemeijer, K., & Ainley, J. (2018). Design of statistics learning environments. In D. Ben-Zvi, K. Makar & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 473–502). Switzerland: Springer Cham.

Ben-Zvi, D., Makar, K., & Garfield, J. (Eds.) (2018). *International handbook of research in statistics education* (Springer international handbooks of education series). Springer.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association, 71* (356), 791–799.

Breiman, L. (2001). Statistical Modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Desai, J., Watson, D., Vincent, W., Mariarosaria, T., & Luciano, F. (2022). The epistemological foundations of data science: a critical review. *Synthese, 200*, 469.

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

Dvir, M., & Ben-Zvi, D. (2018). The role of model comparison in young learners' reasoning with statistical models and modeling. *ZDM—International Journal on Mathematics Education, 50*(7), 1183–1196.

Dvir, M., & Ben-Zvi, D. (2023). Informal statistical models and modeling. *Mathematical Thinking and Learning, 25*(1), 79–99. https://doi.org/10.1080/10986065.2021.1925842

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science (Institute of Mathematical Statistics Monographs)*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316576533

Erickson, T. (2020) Awash in Data: and introduction to data science with CODAP. eeps media. [e-book] (2020). Available at: https://concord.org/awash-in-data

Gafny, R., & Ben-Zvi, D. (2023). Students' articulations of uncertainty about big data in an integrated modeling approach learning environment. *Teaching Statistics, 45*(1), 67–79. https://doi.org/10.1111/test.12330

Hariri, R., Fredericks, H., & Bowers, E. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data, 6*(1), 1–16. https://doi.org/10.1186/s40537-019-0206-3

Hesse, M. B. (1962). *Forces and fields: The concept of action at a distance in the history of physics.* Mineola, NY: Dover.

Holmes, D. (2017). *Big data. A very short introduction.* Oxford: Oxford University Press.

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning, 110*, 457–506. https://doi.org/10.1007/s10994-021-05946-3

Koehrsen, W. (2019). *Thoughts on the two cultures of statistical modeling.* Retrieved August 11, 2023 from Towards Data Science: https://towardsdatascience.com/thoughts-on-the-two-cultures-of-statistical-modeling-72d75a9e06c2

Konold, C., & Miller, C. (2015). TinkerPlots (Version 2.3.1) [Computer software]. University at Massachusetts. Online: http://www.tinkerPlots.com/.

Laney, D. (2001). *3-d data management: controlling data volume, velocity and variety.* META Group Research Note, *6*.

Lee, H., Mojica, G., Thrasher, E., & Baumgartner, P. (2022). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal, 21*(2), 3. https://doi.org/10.52041/serj.v21i2.41

Leonelli, S. (2019). What distinguishes data from models? *European Journal for Philosophy of Science, 9*, 22. https://doi.org/10.1007/s13194-018-0246-0

Manor, H., & Ben-Zvi, D. (2015). Students' emergent articulations of models and modeling in making informal statistical inferences. In A. Zieffler & E. Fry (Eds.), *Reasoning about uncertainty: Learning and teaching informal inferential reasoning* (pp. 57–94). Minneapolis, Minnesota: Catalyst Press.

Manor, H., Ben-Zvi, D., & Aridor, K. (2013). Students' emergent reasoning about uncertainty exploring sampling distributions in an "integrated approach." In J. Garfield (Ed.), *Proceedings of the Eighth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL8)* (pp. 18–33). Minneapolis, MN, USA: University of Minnesota.

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants: A new approach to numeracy* (pp. 95–137). Washington, DC: National Academy of Sciences.

Noll, J., Kazak, S., Zapata-Cardona, L. and Makar, K. (2023), *Introduction to rethinking learners' reasoning with nontraditional data.* Teaching Statistics, 45, S1–S4. https://doi.org/10.1111/test.12350

Plomp, T. (2013). Educational design research: An introduction. In T. Plomp & N. Nieveen (Eds.), *Educational Design Research* (pp. 10–51). Enschede, The Netherlands: SLO.

Rowley, J. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science, 33*(2), 163–180. https://doi.org/10.1177/0165551506070706

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal of Research in Mathematics Education, 26*(2), 114–145.

Simon, M. A., & Tzur, R. (2004). Explicating the role of mathematical tasks in conceptual learning: An elaboration of the hypothetical learning trajectory. *Mathematical Thinking and Learning, 6*(2), 91–104.

Suparman, S., & Maryati, M. (2019). Design of hypothetical learning trajectory (HIDMLT) in mathematics using motif of Anyaman Bambu. *Asian Journal of Assessment in Teaching and Learning, 9*(1), 16–27. https://doi.org/10.37134/ajatel.vol9.no1.2.2019

Tukey, J. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Wild, C. J., Utts, J. M., & Horton, N. J. (2018). What Is Statistics? In D. Ben-Zvi, K. Maker, & J. Garfield (Eds.), *International Handbook of Research in Statistics Education* (pp. 5–36). Springer International Handbooks of Education. Springer, Cham.

Zhang, K., Liu, S., & Xiong, M. (2022). *Changes from classical statistics to modern statistics and data science*. https://doi.org/10.48550/arXiv.2211.03756

Zuboff, S., & Schwandt, K. (2019). *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. Profile Books.

Traditional statistical models in a sea of data: teaching introductory data science

Robert Gould

Dept. of Statistics, UCLA
rgould@stat.ucla.edu

This paper outlines the role of "traditional" statistical models and modeling in K–12 data science education, and compares traditional statistical models with predictive models. I explain why traditional models, with their emphasis on inference within a narrow context, might seem inapplicable to the demands of a modern data science classroom. Nonetheless, these models provide important lessons necessary for students to understand fundamental data science concepts.

Introduction

With the acknowledgement of the importance of data in everyday life comes a growing recognition by many educators for the need to strengthen students' data literacy so that they can work, vote and understand their data-driven world. For example, Lee et al. (2022) describe a growing list of frameworks and guidelines to support high school data literacy in general and data science in particular. Data science courses at the school level (K–12) are one means of teaching important approaches for understanding data.

Data science courses, at least at the K–12 level, may share many of the characteristics of statistics courses, since both fields prepare students to learn about the world through analyzing data. However, data science in the schools differs from statistics in that it exposes students to a greater variety of data and so must require different tools in order to extract meaning from "messy" data. For example, Wise (2020) defines the role of a data scientist as a builder of bridges between unstructured, messy data and interesting although potentially vaguely-posed research questions. Building these bridges requires understanding just how much support a particular data set can provide in addressing a particular question or issue. Erickson suggests that a data science course should make students feel "awash" in data (https://concord.org/awash-in-data/) and should teach "data moves"—strategic approaches and technical procedures such as filtering data—to cope with this sensation (Erickson et al., 2019). This is in contrast to a traditional statistics course, which might provide students with data that include only one or two variables, or only the variables needed to answer a particular problem. A data science course, on the other hand, exposes students to data sets with many variables, some possibly redundant or unnecessary or unexplained. The values of these data are not necessarily numerical (as they tend to be in statistics courses), but might also be text, images, or sounds. The student must also interrogate the data to understand its origin and how it came to sit on the student's computer, and to understand whether the data is capable of answering the student's (or teacher's) questions.

Teaching students to engage with complex data within the mathematics curriculum is challenging, since few mathematics teachers, at least in the United States, have formal preparation in data analysis or computer programming. However, data science rests firmly on a foundation of statistical science, which in turn is supported by mathematics with which teachers may be familiar. The Minerva School invited participants to reflect on the role of *data models and modeling* in education. Mathematical teachers are likely familiar with mathematical modeling. This paper provides insight into the process of statistical models and modeling, shows the relationship to mathematical models, and explains how statistical models support sound reasoning when awash with data.

Because many readers might not be familiar with the role of modeling in statistics, following a brief background to explain the author's interest in this subject, the paper introduces the notion of a "traditional" statistical model as a mathematical model that describes the random variation in observed data. An important take-away of this section is that traditional models require strong constraints on the data that can be modeled, and these restrictions might, at first glance, lead us to believe that traditional models do not belong in a data science course, where one expects students to be exposed to data that do not fit these neat constraints. This is followed by a section that describes the modeling process with these traditional models, and the next section contrasts these traditional approaches with "non traditional" predictive modeling.

The final section explains why, in the author's view, traditional models are important for preparing students for predictive modeling and other techniques that they should learn in a data science course.

Background

This paper is motivated from the author's experience in working on a team that included data scientists, teachers and educational administrators to develop and implement a high school data science course. The author is a statistician and was the principal investigator in a project that culminated in a course called Introduction to Data Science (IDS) (see https://introdatascience.org/). IDS was first taught in 2014 in the Los Angeles Unified School District. IDS is currently taught in 107 districts across the United States and, in the last ten years, has taught over 42,000 students.

IDS was designed with the philosophy that all students need preparation for these engagements with data, and that this preparation may prevent students from being harmed by data while at the same time facilitate their career success. For instance, data can harm us through a loss of privacy that may occur when data sets are linked together and personal facts revealed, or may harm a student's sense of agency or identity, if algorithms classify students in ways that conflict with their cultural or personal identities. The IDS curriculum was developed on the notion that the bedrock of data science education is statistical reasoning. The primary influences on the curriculum were the American Statistical Association's GAISE K–12 report (Franklin et al., 2005) and the U.S. Common Core standards for mathematics (CCSSI 2010), with some influence from the CSTA principles (Seehorn et al., 2011). IDS extends the traditional statistics course in several ways. First, it goes beyond situations in which data have been collected solely for the purposes of statistical inference and includes data collected through sensors and a paradigm called participatory sensing, in which students use mobile devices to collect a rich set of multivariate data. Second, IDS includes a wider variety of data types than traditionally encountered in a statistics classroom. These include text, images, times and dates, and locations. Third, it relies strongly on computing to assist preparing the data for analysis and the analysis itself. Finally, it includes predictive modeling, a methodology that is not included in statistics curricula at the school level (at least not formally).

An important goal in the design of IDS was to prepare students to reason with and learn from the types of data that they were likely to encounter in their everyday lives, and not the "classroom" data that many statistics classes provide. These "everyday" data include data from open data portals, sensors, and mobile devices. The goal is that students learn to pose their own investigatory questions on topics of interest to them, and to evaluate whether available data is suitable to answer their questions. At a more detailed level, though, what, precisely, do students learn in a data science course?

Traditional statistical models and mathematical models

IDS focuses on a modeling process called the "Data Cycle". The Data Cycle is indirectly based on the "Problem, Plan, Data, Analysis, Conclusion" Cycle, or PPDAC, (Wild & Pfannkuch 1999) and directly on the GAISE statistical investigation process. Consistent with Schulte (2024, **in this volume on page 61**), this cycle might also be described as a problem-solving process. This cycle contains four stages: Ask Questions, Consider Data, Analyze Data, and Interpret Data. Students need experience and practice in all four phases. For example, the "question" phase is tackled early in IDS, since experience has shown that posing productive statistical investigative questions is challenging for both students and teachers (Gould, Bargagliotti, and Johnson 2017; Frischemeier and Biehler 2017; Frischemeier and Leavy 2020; Bar 2022).

The Data Cycle and the PPDAC share many similarities with descriptions of mathematical modeling. For example, the Common Core Mathematical standards (a set of standards adopted, with or without modifications, by many of the states in the U.S.) describe a mathematical modeling cycle with four phases. The cycle starts with a formulation of a problem (similar to "Ask Questions"), moves to computations ("Analyze data"), then to interpretations ("Interpret Data"), and then to validation (some of which is done in the "Consider Data" step). At that point, depending on the outcome of the validation procedure, the modeler either re-formulates the problem or reports their findings (CCSSI, 2010). Given these similarities, what is gained by carving out a distinction from mathematical models for statistical models and modeling? In a word: uncertainty.

A statistical model, in its most traditional sense, consists of two components: signal and noise. (For example, see Chatfield 1995). Or, if you prefer, trend and variation, or a deterministic component and a stochastic component. That second component, the variation component, describes uncertainty and distinguishes a statistical model from a mathematical model. The "signal" component by itself is represented by a mathematical function that describes a general relation, typically between the mean value of our response variable and observable characteristics (i.e., input). For example, the relationship between the mean salary paid to an institution's employees might

be a linear combination of their rank, years of experience, the year they began employment, and their sex. The trend component should consist of all factors that contribute causally to the response variable.

The noise component tells us how actual observations vary about the signal. This is an important feature of statistical models and the feature that distinguishes them from mathematical models. A physicist might derive a mathematical model that predicts the precise distance traveled by a tennis ball hit at a particular angle with a given force, but only when the model includes random deviations from that predicted distance is it statistical. Bielik (2024, **in this volume on page 33**) describes tools to help students to use data to develop and investigate the trend component within the context of ocean acidification. While this exploration is data-driven and complex, the trend component alone is not a statistical model, but can be made statistical by including a stochastic component.

The stochastic component specifies the probability distribution from which these deviations arise, and also specifies whether observations are associated or independent of each other. It is because of the stochastic component that statisticians caution that when we view the world through data, we do not see the world as it actually is, but instead through the distortions of this noise (Wild, et al., 2011). Put differently, the stochastic component tells us that our data collection could have turned out differently, and reminds us that we see just one of many possible outcomes. Without the stochastic component, a model of data may communicate a misleading degree of certainty and precision.

Statisticians sometimes speak of statistical models as modeling the "data generating process." In this view, a successful statistical model can be viewed as an algorithm that can generate a simulated data set that is indistinguishable, in all important characteristics, from the actual data. For this reason, these traditional statistical models can be viewed as data models: abstract descriptions of how the data came into existence. For example, Buja et al. (2009) describe hypothesis testing procedures in which a proposed statistical model is used to generate multiple versions of the data. Each of these versions is summarized with an appropriate graphic which can then be compared to the same graphical summary of the original data. If the original graph is distinguishable from the graphs generated by the model, then the model of the data generation process is wrong.

Statistical models are traditionally communicated using mathematical notation. Consider a model of reality proposed by Vitruvius, the ancient Roman architect. He proposed that a person's arm span (the distance from finger tip of the longest finger of the left hand to the longest finger of the right when arms are held out horizontal to the ground) was equal to their height (a model famously visualized by Leonardo da Vinci as the *Vitruvian Man*). This model can be stated as a linear model:

$$\text{armspan} = 0 + 1 \times \text{height}$$

This is a mathematical model, and while it might have been motivated by observation, it is not based on data. Suppose we were to collect data on a sample of humans and plot their arm span lengths against their heights. This model predicts we will see a perfect line with intercept 0 and slope 1. Instead, we see something like Figure 1:

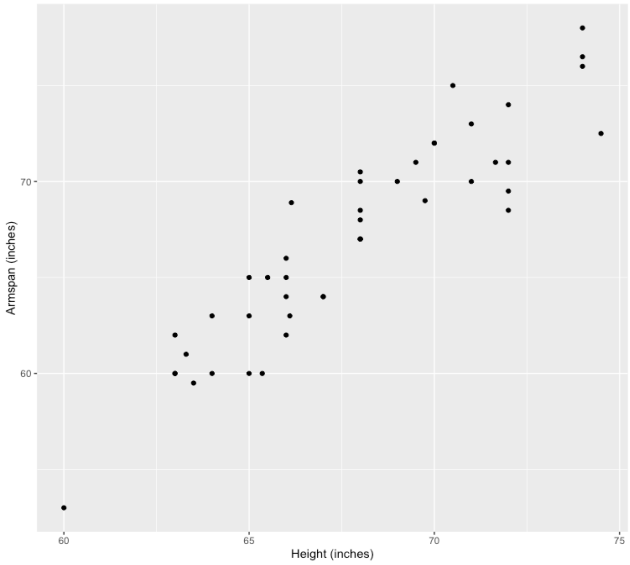


Figure 1: Armspans and heights (in inches) as reported by students at the University of California, Los Angeles.

Rather than a line, we see a trend which is generally linear but with great variation about the trend. The variation may have several sources: environmental, measurement errors, etc. A statistical model for this relationship includes this variation about the trend. One possibility is:

$$\text{armspan}_i = \beta_0 + \beta_1 \text{height}_i + \epsilon_i$$

The subscripts i denote the individuals in the data set. More generally:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where

$$\epsilon_i \sim \text{iid} N(o, \sigma)$$

The trend component of the model is $\beta_0 + \beta_1 x_i$. The beta parameters represent numbers that are unknown and will be estimated from the data. The x variable represents an independent attribute that is purportedly related to the attribute represented by the y variable. The stochastic part is represented by the ϵ (*epsilon*) term and the description of its probability density. This model tells us that deviations from the trend (which are represented by ϵ) are random as determined by a Gaussian distribution with

mean 0 and standard deviation σ (*sigma*). The parameter σ is a number which is unknown but can be estimated from the data. The letters "iid" stand for "independent and identically distributed," which tells us that each deviation is independent of all of the others, and each of them is a random value drawn from the same (Gaussian or "Normal") distribution. This claim about the distribution is an assumption of the model and not derived from an examination of the data. However, it is an assumption which can be verified (or refuted) through considering the data.

The trend component, a mathematical model, tells us something about the fundamental relationship between *x* and *y*. If Vitruvius is correct, then our intercept will equal 0 and the slope 1. The trend component *with* the stochastic component tells us what our data actually look like. The stochastic component also implicitly situates the data in a context in which these observed data are just one of infinitely many possible manifestations of data sets we might have seen. The analysis phase of the modeling cycle will attempt to infer the trend from the cloud of points in Figure 1 so that we can evaluate Vitruvius's theory.

Other traditional statistical models include, to name a few, time-series models (which often specify a correlation between observations close in time), hierarchical linear models (sometimes called nested models or mixed models, which specify correlations between observations within the same level), spatial models (which specify correlations between observations nearby in space), survival models, and logistic models.

Having clarified our definition of a traditional statistical model, we now consider the process through which a model is fit to the data.

Modeling with traditional models

Once a traditional model is proposed as a candidate for the data description process and also as a potentially useful model for addressing posed investigative questions, the analyst then engages in a process in which diagnostic tools are applied to investigate and to minimize, by adjusting the model, the gap between the model and the data. Ideally, this process is carried out on several candidate models and the results compared (a principle D'Agostino McGowan, Peng and Hicks (2022) refer to as "exhaustiveness"). In many statistics classrooms this iteration between a class of models and data might be identified as the "modeling process." The IDS curriculum prefers to more expansively define the modeling process as part of the Data Cycle, which includes this traditional process as an activity within the "analyze" phase.

Ideally, this analysis process is guided by some knowledge of the context, which tells the analyst which variables are viable components of the trend and which are not, and what, precisely, the functional relationship might be. Without this contextual knowledge—if, for example, the variables were provided to the analysts with non-descriptive names such as *x1*, *x2*, and *y*—then discovering the "true" model can be considerably challenging, if not impossible.

For a toy example that nonetheless uses real data (the trees data in the *datasets* package in R, (R Core Team 2022), consider this analysis in which we ask whether we can model the volume of a tree as a function of its height and diameter. A linear regression model is a reasonable place to begin:[1]

$$\hat{V} = \hat{\beta}_0 + \hat{\beta}_1 H + \hat{\beta}_2 D$$

V represents volume, *D* diameter and *H* height. The "hat," or carat, over the parameters and variables indicates that this is a value estimated from the data. Guided by the application of various diagnostic tools (for example, examining residual plots), a reasonable modeling process that considers different forms for the trend leads to this model:

$$\hat{V} = -72.5 + 0.3H + 0.2D^2$$

Distressingly, this model provides a 95% confidence interval for the intercept as (–41, –14); in real life, we would expect trees with 0 height and diameter to also have 0 volume. Clearly this model falls short (and it falls short according to other diagnostic methods as well).

We can make much better progress armed with a crude theory: trees are like cylinders. The volume of a cylinder is given by

$$V = \pi(D/2)^2 H$$

That is, volume is related to a *product* of diameter and height terms rather than a sum. We therefore recast that volume formula to one that finds the log of the volume:

$$\log V = -.24 + 2 \log D + 1 \log H$$

Fitting this model produces relatively clean diagnostics and a more believable final model:

$$\log V = 6.6 + 2.0 \log D + 1.1 \log H$$

Confidence intervals for the slopes are in agreement with our cylinder theory, although the model is mathematically

1 The models here omit the subscripts that were present in the models in the previous section. The reason is that the models discussed in this section are models of the general trend component whereas the previous section's models described individual observations with their included random variability.

silent on the issues of trees with 0 diameter and height (which is also an improvement over the previous models).

Note that we focused on the trend component of the model to estimate the parameters. The modeling process also produces estimates of the variability about the estimated trend. The estimated variability is needed in order to perform inference, such as estimating confidence intervals for the slopes and intercept.

In the above example, the system being modeled was fairly simple in that there was no need to understand causal relations between the predictor and response variables. Other situations include causal questions: why are crime rates changing? What affect do certain behaviors have on the risk of catching COVID-19? In these situations, particularly when the data are not from designed and controlled experiments, we often cannot be sure whether we have all of the necessary explanatory variables in the model. Kronmal (1993) provides an entertaining artificial example in which the association between the number of storks in a town and the number of babies born is strong and positive, but disappears entirely when the number of women is included (Snow 2020). Real-life, complex systems may include hundreds of variables and these variables might not be included in a model (or even known). These situations are particularly dangerous for modelers, since omitting important variables from models can lead to substantial bias; associations may be reported as positive when they are actually negative (or vice versa) or are in truth unrelated. On the other hand, including unimportant variables can lead to overfitting (which can lead to large prediction error or other unreliable descriptions of phenomena).

Four things to note:

1. Traditional statistical models are focused on understanding real-world phenomena, and designed to allow for the statistical inference needed to distinguish "true" patterns from random ones.
2. The restrictions required to achieve sound inference are strict and require heavy oversight during data collection or, at the very least, detailed knowledge of the data collection procedure. As a result, data are best collected by experts in the substantive field who collaborate with the data analyst.
3. The data are assumed to be clean and appropriately organized for the intended analysis. In other words, no more data moves are required.
4. The modeler relies on context to shape the model and to understand which variables must be considered and which need not be considered. But there is no guarantee that all necessary variables will be included (e.g., there is no guarantee that the model is free from bias) or that some unrelated variables will be included (resulting in an inability to properly generalize).

Traditional statistical models and introductory data science

Do traditional statistical models and modeling belong in a high school data science course? How do they help students build bridges between research questions and messy, unstructured data? How do they prepare students for studying machine learning, a common type of modeling in data science? To begin, let's examine predictive modeling, a core data science activity.

Predictive models and modeling

I've taken care to use the adjective "traditional" in this discussion, because in the last few decades, another class of models have risen in importance. These traditional models we've described so far might be considered "inferential" in that we're using the model to infer a broad, general pattern based on our sample. For example, we wish to know the relation between arm span and height for *all* humans, not just those in this particular data set, and we wish to confirm whether this slope is equal to 1 were we to see all humans' data. But there's another class of problems for which data are modeled and which are of great importance to data science, namely, predictive problems. A simple, but perhaps uninteresting, example of a predictive model might be that if I'm told a person's height, I want to be able to predict their arm span with a useful level of precision. More interesting examples occur in the contexts of online shopping (predicting the amount a customer will spend or perhaps specifically which items the customer will seek), medical diagnosis, and weather and climate forecasting. The field of machine learning is essentially the study of predictive modeling.

In 2001, statistician Leo Breiman famously described a culture clash between inferential models and predictive models, claiming, with good reason, that the predominant statistics culture wasn't paying sufficient attention to the important area of predictive modeling (Breiman 2001). These predictive models differ from the traditional models in that they are not always communicated via mathematical notation; instead, they may be algorithmic, or consist of ensembles of algorithms that each "vote" on a prediction with the prediction receiving the most votes declared the winner. These ensembles are complex and not transparent; it can be difficult to explain precisely which factors were relevant in the models' output.

An example of an algorithmic predictive model that is accessible to high school students is a decision tree. Podworny and Frischemeier (**in this volume on page 15**) describe the use of decision trees in K–12 data science education, and the IDS curriculum also includes a unit on decision trees. These trees, which Breiman formalized as Classification and Regression Trees (CART)

(Breiman 1984), are perhaps the foundational example of predictive modeling. As Podworny and Frischemeier show, trees provide a way for students to understand misclassification rates, overfitting and other fundamental concepts. They are essential for understanding more advanced approaches that consist of ensembles of algorithms, such as random forests and boosting and bagging techniques.

Figure 2 provides an example of a decision tree based on data used in IDS. In an introductory exercise in classification, students are asked to determine rules for sorting athletes into two groups: soccer ("football" in most of the world) or (American) football. The data provides the athletes' age, weight and height. The decision tree shown is produced algorithmically using the *rpart* package provided in R (Therneau & Atkinson 2022). For example, this model tells us that athletes that weigh less than 200 pounds should be classified as soccer players (US Men's National Team, USMNT), that 38% of the sample were classified this way for this reason, and that there was 100% success rate. On the other hand, if an athlete weighs 213 or more pounds, then they should be classified as an American football (NFL) player. 56% of the sample was so classified, with a 96% success rate.

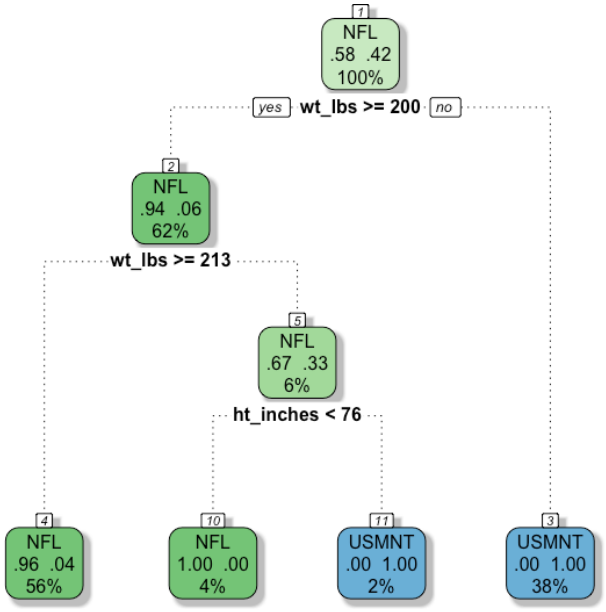


Figure 2: Tree for classifying athletes as soccer players (USMNT) or American football players (NFL). The bottom line lists the proportion of the sample assigned to that node. The middle line in each box shows the proportion of players in that node from the NFL and USMNT groups. The top line indicates which type of player is in the majority.

Traditional models support data science

Given that predictive models are of central importance to data science, and that data science courses prepare students to work with data that may not conform to the constraints required of traditional statistical models, why should data science study traditional statistical models? The reason traditional models are still relevant is that the lessons learned from traditional models and modeling inform almost every aspect of data science, particularly data description, data visualization, and predictive modeling.

Let's begin with data description. Data description is an important analytic skill, particularly in contexts in which the data lack a describable data-generating mechanism or when the data are multi-dimensional and require a high level summary. For example, IDS students collect data on their daily routines using their mobile devices. One such data collection campaign tracks "snacks": food eaten between meals. The resulting data are not randomly sampled (data are collected every single time a student snacks, over the course of a few days) and may have strong measurement biases, and so determining the stochastic component of a traditional model as described above is not possible, despite the presence of variability. The data are highly multivariate and include text data (names of foods), categorical values (health ratings, reasons for eating), time, date, location, and numerical values (how many people were you with when eating?) Given these conditions, it might not be possible to develop a useful inferential model; a more useful approach to these data is to provide a thorough description of important features of the data.

Measures of center and spread are the primary tools for univariate descriptions. The notion of the mean is complex and abstract (for examples, see Garfield 2007). When taking a mean, we make a tacit assumption that the observations come from the same population or, more informally, the same group. One could calculate the mean circumference of apples and oranges, but one really should calculate these means separately so as not to merge two distinct groups. The mean is a useful summary when the data generating model is this traditional statistical model:

$$y_i = \mu + \epsilon_i$$

where the random deviations, represented by epsilon, come from some symmetrically-shaped probability distribution. The implication of this model is that the data are drawn from the same population and are different measurements of something whose true value is represented by μ. As noted above, a traditional statistical model sits at the heart of the notion of a mean. It turns out that our simple calculation of a mean of a variable comes laden with assumptions.

Introductory students needn't learn about this mathematical model explicitly. But they should know that this simple act of calculating a mean is actually an act of fitting a model and, like all models, the fit may be useful or useless. When calculating a mean, the model assumes that the observations belong together; they come from the same population. This simple model is of fundamental importance today, as the role of aggregation, which was controversial when Quetelet proposed it, is still debated, perhaps particularly so in the context of diversity, equity and inclusion.[2] For example, Zelnick (2021) discusses the issues surrounding the inclusion of a racial variable in models to predict eligibility for kidney transplants. Including the racial variable results in a model in which calculations do *not* aggregate across race. Yet excluding the variable means we ignore racial categories (as recorded in the data). Kaufman and Cooper (2001) discuss the use of racial variables in epidemiological work, and Thornton et al. (2022) discuss the challenges of using gender and sex as categories.

When engaging in descriptive data practices, students might naturally focus on small differences in means between groups. The lesson that "what we see is not exactly the way the world is" is the first step in developing caution about the stories one tells from data, and, if handled carefully, begins to develop what Makar and Rubin (2009) define as "informal statistical inference." Gafny and Ben-Zvi (**in this volume on page 69**) study how students—graduate students in this case—express uncertainty when exploring "big data" and describe how this can be used to develop this form of inferential reasoning. Developing students' informal inferential reasoning is important, because when the data generating process does not conform to that required by traditional models, it may be the only form of inferential reasoning available. Even when traditional models might be potentially applicable, Kreuter (2017) reminds us that much that happens in the data-generating process is not captured by traditional models. For example, "…people get sick, are on vacation, or don't want to participate" in your randomly sampled survey (p. 420).

Next, consider data visualization, which is perhaps the first analysis tool that students learn in statistics and data science. Indeed, entire data science lessons are built around interpreting visualizations. The New York Times provides the resource *What's Going on in This Graph* (WGOITG) (https://www.nytimes.com/column/whats-going-on-in-this-graph) to help students critically analyze multi-dimensional data displays. Although perhaps best suited for beginners, WGOITG offers useful lessons for more advanced students as well (Arnold et al., 2022). For example, the graphic for January 11, 2023 showed, for three-plus flu seasons starting with the 2019–20 year and ending a few months into the 2022–23 season, time-lines for the percent of weekly doctor and hospital visits for respiratory illness by day of the year. Three variables are displayed: year, percent, and day of the year. A "pre-pandemic average" line is also provided. The prompt asks "What do you notice and wonder about the intensity, peak and duration of respiratory illnesses in the U.S.?" One might notice, for example, that in the past, respiratory illnesses peaked in January; but for the 2022–23 flu season the number of illnesses is higher than in the past and the increasing trend line is considerably steeper than in the past.

Traditional statistical modeling teaches us of the dangers of omitting important variables from our model. Data visualizations are limited by the print media to a small number of variables (often 2 or 3), and so are particularly susceptible to this form of bias. Students must therefore learn that when they interpret data visualizations, they must also think about what is missing from the graphic. The WGOITG visualization invites us to conclude that the 2022–23 season is remarkable, with respect to respiratory illnesses. Yet the graph shows only three variables and surely there are omitted variables. Traditional modeling forces us to wonder what could be missing that might alter the story. The total number of visits? Age? Climate variables? The inclusion of years prior to 2019–20? (The pre-pandemic average line shows us the general trend for earlier years, but not the variability in that trend.)

Finally, we return to predictive modeling, which provides an interesting contrast with traditional statistical modeling as described above. Traditional modeling is challenging because the goal is to discover the "truth," and truth is famously elusive. In predictive modeling, however, the model that best predicts future observations is the best. There is no need to worry about whether or not the model is true as long as its predictions are usefully precise. If our decision tree usefully classifies visitors to our shopping website as "buyers" or "non-buyers," that may be all we need, and we may not be concerned with which variables played a role in this classification.

Despite such differences, modeling with traditional models provides important lessons for predictive models. The apparent simplicity of a decision tree hides uncertainty and variation and so, at first glance, a decision tree might not seem to be a statistical model, as defined in this paper. And yet, decision trees are the product of uncertainty and variation; other manifestations of data drawn from the same population may produce slightly (or even radically) different trees. Students experienced with traditional models will know that, where data are

2 Raper (2017) provides an historical overview that illustrates the conceptual challenge of using the mean as a summary statistic. Gigerenzer, et al. (1989) Chapter 2 provides an overview of Quetelt's *l'homme moyen* and arguments for and against this concept.

concerned, what we see is not exactly the way the world is. The misclassification rates provided at the terminal nodes of the tree are estimates of misclassification probabilities, and students with experience in traditional models should understand that, as estimates, they model uncertainty.

Traditional statistical models provide an explicit account of variation and so are accessible to those with mathematical (but not statistical) backgrounds. But these traditional statistical models also provide an important conceptual bridge to the predictive models used in data science, where the uncertainty may be implicit, but is always present.

Conclusion

The 2022 Minerva School invited us to consider the role of data models and modeling in data science education. As a statistics educator who has been involved in high school data science education, the importance of building data science learning on a strong foundation of statistical reasoning cannot be underestimated (Gould 2021). This paper has provided a description of traditional statistical models as consisting of a trend component and a stochastic component. The latter component is viewed as a model of the randomness inherent in the data-collection process and is necessary for quantifying uncertainty in generalizations beyond the data at hand. This stochastic component can be viewed as a set of conditions required for inference; as such, it is quite restrictive. Because of this restrictiveness, one might believe that traditional modeling is irrelevant in a data science classroom in which students consider data from non-random samples or that violate these strict conditions in other ways.

The last section of this paper was intended to convince the reader that, in fact, traditional models belong in the data science classroom. The lessons learned from studying traditional models—that context matters, that what we learn about the world when viewed through data is not exactly the way the world actually is, and that variability and therefore uncertainty are an inherent part of data analysis—influence common data-science-classroom lessons involving data description, data visualization, and predictive modeling.

It's true that traditional models were developed to enable a very particular type of analysis, namely generalization from small samples to large populations under very rigid data collection protocols. While this type of analysis is extremely powerful and propels much of scientific progress, given the wide variety of data that now surrounds us, this approach can seem quite limited, particularly to younger students for whom data collected for scientific studies might feel quite remote. Still, the fact that the data did *not* come from a random sample or a well-designed experiment does not give license to plow forward at full speed. If anything, particular caution is warranted. These "traditional" models show us what caution looks like.

A final benefit of statistical traditional models is that they provide a familiar bridge to teachers of mathematics. Both the trend component and the stochastic component are mathematical objects, but the struggle to use these mathematical objects to model real-world phenomena as viewed through the rippled glass of data is a core statistical practice. Mathematics teachers might find statistical modeling with traditional models more familiar than algorithmic/predictive modeling approaches. The core practices and concepts applied in traditional modeling can then provide a bridge towards algorithmic modeling, which does not always result in a closed-form mathematical expression.

The papers presented at the Minerva School illustrate the variety of approaches and conceptions, as well as desired learning outcomes, situated around data models and modeling. This variability is no doubt induced, in part, by the still-emergent nature of the field of data science itself. This paper has attempted to describe a vision of data science that places emphasis on analyzing and modeling data. In this vision, traditional statistical models, no matter how limiting they may appear when viewed through the demands of modern data, provide important lessons we must consider when designing curricula to support data science education.

References

Arnold, P., Bargagliotti, A., Franklin, C., & Gould, R. (2022). Bringing Complex Data Into the Classroom. *Harvard Data Science Review*, 4(3). https://doi.org/10.1162/99608f92.4ec90534

Bar, C. (2022). Dataset-driven instruction in the biology classroom. Paper presented at the Minerva School 2022: *Reasoning with data models and modeling in the big data era.*

Bielik, T (2024). **"Supporting students' modeling and data practices by engaging with digital tools" in this volume on page 33.**

Breiman, L., (1984) *Classification and Regression Trees* 1st Ed., Routledge, New York.

Breiman, L., (2001) Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.* 16(3): 199–231, August 20010. DOI: 10.1214/ss/1009213726

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E., Swayne, D.F., Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Phil. Trans. R. Soc., A* (2009) 367, 4361–4383 doi: 10.1098/rsta.2009.0120

Chatfield, C., (1995), *Problem Solving: A statistician's guide,* 2nd ed., pp. 26–28. Chapman & Hall/CRC, Boca Raton, London, New York, Washington, D.C., 1995.

Common Core State Standards Initiative (CCSSI). (2010). Common Core State Standards for Mathematics. Washington, DC: National Governors Association Center for Best Practices and the Council of chief State School Officers.

D'Agostino McGowan, L., Peng, R.D., Hicks, S.C. (2022) Design Principles for Data Analysis. *Journal of Computational and Graphical Statistics*, 00(0)1–8.

Erickson, T., Wilkerson, M., Finzer, W., & Reichsman, F. (2019). Data Moves. *Technology Innovations in Statistics Education*, 12(1). http://dx.doi.org/10.5070/T5121038001 Retrieved from https://escholarship.org/uc/item/0mg8m7g6

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., Scheaffer, R. (2005). Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K–12 Curriculum Framework. The American Statistical Association. https://www.amstat.org/docs/default-source/amstat-documents/gaise-prek-12_full.pdf

Frischemeier, D., and Biehler, R., (2017). Stepwise development of statistical literacy and thinking in a statistics course for elementary preservice teachiers. In: *T. Dolley & G Guedet (Eds.), Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education, 756–763*. Dublin, Ireland: DCU Institute and ERME, 2018.

Frischemeier, D. and Leavy, A. (2020). Improving the quality of statistical questions posed for group comparison situations. *Teaching Statistics*, 42:58–65. DOI:10.1111/test.12222

Gafny, R. and Ben-Zvi, D. (2024). **"Reimagining data education: Bridging between classical statistics and data science" in this volume on page 69.**

Garfield, J., (2007). How Students Learn Statistics Revisited: a Current Review of Research on Teaching and Learning Statistics. *International Statistical Review*, 2007. 75(3). 372–396. https://www.jstor.org/stable/41509878

Gigerenzer, G., Switjink, Z., Porter, T., Daston, LO., Beatty, J., and Kruger, L., (1989*), The Empire of Chance*, Cambridge University Press, New York.

Gould, R., Toward data-scientific thinking, *Teaching Statistics* 43 (2021), S11–S22. https://doi.org/10.1111/test.12267

Gould, R., Bargagliotti, A., Johnson, T. (2017). An Analysis of Secondary Teachers' Reasoning with Participatory Sensing Data. *Statistics Education Research Journal*, 16(2) November 2017.

Kaufman, J.S., and Cooper, R. S. (2001). Commentary: considerations for Use of Racial/Ethnical Classification in Etiologic Research. *America Journal of Epidemiology*, 154(4), August 15, 2001.

Kreuter, F. (2017), Inference from Big Data: A Cross-Disciplineary Endeavor, in *International Handbook of Research in Statistics Education*, Ben-Zvi, D., Makar, K., and Garfield, J., (eds), Springer International Publishing.

Kronmal, Richard A. (1993) Spurious Correlation and the Fallacy of the Ratio Standard Revisited. *Journal of the Royal Statistical Society. Series A*, Vol. 156, No. 3, 379–392

Lee, H., Mojica, G., Thrasher, E., & Baumgartner, P. (2022). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal*, 21(2), 3

Makar, K., and Rubin, A., (2009) A Framework for Thinking About Informal Statistical Inference, *Statistics Education Research Journal* 8(1): 82–105.

Podworny, S. and Frischemeier, D. (2024). **"Young learners' perspectives on the concept of data as a model: what are data and what are they used for?" in this volume on page 15.**

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ .datasets::trees

Raper, S. (2017). The shock of the mean. *Significance Magazine,* December 2017, 12–16. Royal Statistical Society.

Schulte, C. (2024). **"From representation to transformation: rethinking modeling in computer science education" in this volume on page 61.**

Seehorn, D., Carey, S., Fuschetto, B., Lee, I., Moix, D., O'Grady-Cunniff, D., Owens, B., Stephenson, C., Verno, A. (2011). CSTA K–12 Computer Science Standards: Revised 2011. Computer Science Teachers Association and the Association for Computing Machinery, Inc., https://csteachers.org/teacherstandards/

Snow G. (2020). *TeachingDemos: Demonstrations for Teaching and Learning*. R package version 2.12, https://CRAN.R-project.org/package=TeachingDemos.

Therneau T, Atkinson B (2022). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1.19, https://CRAN.R-project.org/package=rpart.

Thornton, S., Roy, D., Parry, S., LaLonde, D., Martinez, W., Elli, R., Corliss, D (2022). Towards statistical best practices for gender and sex data. *Significance Magazine*, February 2022. The Royal Statistical Society

Wild, C., Pfannkuch, M., (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, 67(1999), 223–265.

Wild, C. Pfannkuch, M., Regan, M., Horton, N.J. (2011). Towards more accessible conceptions of statistical inference. *J.R. Statisti. Soc. A.* (2011), 174, Part 2, 247–295.

Wise, A.F. (2020). Educating Data Scientists and Data Literate Citizens for a New Generation of Data. *Journal of the Learning Sciences* (2020), 29:1, 165–181.

Zelnick LR, Leca N, Young B, Bansal N. Association of the Estimated Glomerular Filtration Rate With vs Without a Coefficient for Race With Time to Eligibility for Kidney Transplant. *JAMA Netw Open.* 2021;4(1):e2034004. doi:10.1001/jamanetworkopen.2020.34004

What do citizens need to know about real-world statistical models and the teaching of data modeling

Iddo Gal

Department of Human Services
University of Haifa, Israel
iddo@research.haifa.ac.il

This chapter aims to contribute to the improvement of pedagogies and teaching that can promote learners' ability to act as critical consumers of model-based statistical messages in the real world. The chapter advocates for the need to employ pedagogies that can maximize skill transfer and hence adopt an external view on the real-world demands for model-related knowledge in different life contexts. The chapter focuses on three such contexts, involving media interpretation, service consumption by service users, and workplace environments. It sketches an innovative framework that involves three general families or types of models that citizens and workers have to deal with, i.e., aggregation models, prediction models, and simulation models, as well as three frameworks that all these models usually include, encompassing a conceptual framework, methodological framework, and computational and reporting framework. The chapter discusses implications of these ideas for needed pedagogies that improve learners' ability to act in a critical way when encountering data-based models in real-world contexts, and for future instructional approaches and research directions.

Introduction

This conceptual chapter is inspired by the general aim of the book, which continues the goals of the Minerva School held in 2022, to broaden the perspectives on teaching data modeling and enable learners to become functional and productive citizens of tomorrow. Specifically, the chapter aims to contribute towards the third of the three big themes that motivate this book: "What theories and pedagogies are needed to promote and study reasoning with data models and data modeling?"

The chapter presents general ideas about real-world statistical models and teaching about data-based models that can promote learners' ability to act as critical consumers of model-based statistical messages in the real world (Gal, 2002; GAISE, 2016). The proposed ideas are relevant both to educators working in school systems, especially at the high-school and middle school levels (where statistics is usually subsumed within mathematics education), and to lecturers working in academic institutions (where statistics is usually taught as an independent subject across a wide range of academic departments and subjects, at times in association with the teaching of research methods). In both of these learning contexts, the interest in students' modeling abilities is of growing interest (e.g., Cevikbas et al., 2022). That said, this chapter may be of relevance for educators working in other areas as well, since teaching about data-based models is also addressed in several STEM domains, e.g., in computer science or information technology, environmental science, and biology.

The chapter is organized in four key sections. **"Perspectives on teaching about models and modeling, and skill transfer" on page 92** points to the need to worry about the connection between how models are introduced in class and the real-world demands for knowledge about models, and about skill transfer in this regard. **"About real-world contexts for data-based models" on page 93** distinguishes three separate but related life contexts in which model-related knowledge is needed: media interpretation, service consumption, and workplace processes. **"Characteristics of statistical models in the real world" on page 94** aims to advance the understanding of key issues that should be addressed in instruction about data-based models by presenting three general families of models that citizens (and workers) have to deal with, classified in terms of their use or purpose, followed by three frameworks that all models arguably include and that learners should be made aware of. **"Discussion: implementation, critical thinking, research directions" on page 97** discusses implications of these ideas for learners' ability to act in a critical way when encountering data-based models in real-world contexts, and for curriculum design and future research.

Perspectives on teaching about models and modeling, and skill transfer

Modeling in mathematics education vs. statistics education

In general, a model is a representation (but also simplification) of reality, which aims to capture and reflect key elements in that reality, as well as the relationships or influences between them (Doerr & English, 2003). In mathematics education (K–12 level), the need to lead students through the well-known "mathematical modeling cycle" (Blum & Borromeo Ferri, 2009) is taken as an axiom, and this cycle includes a discussion of (mathematical) models. Illustrating this line of thinking is the established and influential line of work on "model-eliciting tasks" by Lesh and colleagues (e.g., Lesh and Lehrer, 2003). Gravemeijer & Doorman (1999) have argued that the use of models from a mathematics education perspective aims, among other things, to shift learners from viewing a model of a situated activity to a model that is for mathematical reasoning. A recent systematic literature review confirmed the emphasis on mathematical modeling competencies and on students' ability to mathematize and create relevant models within given situations (Cevikbas et al., 2021).

Thus, the prevailing thinking in mathematics education appears to be driven by an *internal* view of "models-for-modeling," i.e., it reflects a desire to engage learners with the conceptual building blocks and tools provided by the discipline (mathematics), and to help them see how these tools can be useful for solving real-world problems. It is important to emphasize that that the mathematical modeling cycle aims to enable students to engage with solving *any* type of problem via a transfer between an everyday reality and a mathematical analysis. This means that students may analyze any type of quantitative information that pertains to a "real" situation, yet such information is neither limited to statistical models nor to "data" in the statistical sense (see Doerr & English, 2003), which is the focus of this book and of this chapter.

When it comes to teaching for understanding *statistical* models, things become more complex, both for teachers and learners alike, and learning goals have to be reconsidered and expanded. As Gould argues (2024, **in this volume on page 81**), a traditional statistical model involves two components: signal and noise (or trend and variation, or deterministic and stochastic components). According to Gould, the signal is a component consisting of all the factors that contribute causally to the response variable, and is represented by a mathematical function that describes the general relation between the mean value of our response variable and observable characteristics. The noise component is the distinguishing feature of statistical models, as it tells us how actual observations vary randomly about the signal.

Additional ideas are associated with the differentiation between mathematical and statistical models, for example:

- Shmueli (2010) and other scholars in statistics education (e.g., Garfield et al., 2008) explain that statistical modeling employs *data*, which may *change* over time and differ in terms of its nature (e.g., how it was collected, what kinds of instruments or definitions were used), measurement properties, volume or amount, or completeness.
- From an educational perspective, learning about seemingly simple statistical models (e.g., average) or more advanced ones (correlation or regression), requires understanding of how *variability in the data* can be modelled by a statistic (Lehrer & English, 2018).
- Moreover, the above implies that learners who engage with statistical models have to grapple with new or more complex ideas that also require the critical evaluation of models, such as regarding *goodness of fit* of models to data, or with the degree of *error* (or variance explained) inherent in any model, or in a prediction based on the results of a model.
- Finally, numerous scholars have emphasized the centrality of the context from which data emerge, and its importance for making sense of results of a modeling process, as a distinguishing feature of statistical modeling (e.g., Pfannkuch et al, 2018).

Accordingly, scholars in statistics education have highlighted the importance of acknowledging and focusing on the *statistical modeling cycle*, which differs in some key aspects from the mathematical modeling cycle, and has specific educational implications (e.g., Wild and Pfannkuch, 1999; Son et al., 2021; Zieffler et al., 2021). In the teaching of statistics within school mathematics, modeling is usually approached by having learners analyze some raw data (as is illustrated by some of the other chapters in this book). The same holds for college level instruction in introductory statistics, as evident in the recommendations in the influential GAISE report (2016), which emphasizes the importance of having students analyze "real data." Indeed, learning to analyze and model data with statistical tools can accomplish key curricular goals set for college and high-school.

However, this arguably has only partial connection to the nature of the statistical models that citizens actually encounter in central life contexts, as discussed in subsequent sections. Further, as Gal (2002) argues when discussing the conceptualization of statistical literacy, in key real-world contexts, regular citizens virtually never analyze any raw data (nor build models!), but rather have

to act as "smart consumers" of already digested information. This includes messages about statistical models and data-based modeling. Before discussing these issues in more detail, we have to acknowledge some concerns about skill transfer, discussed below.

The need to think of skill transfer, and seek an "external" lens onto models

The approaches to models and modeling in mathematics and statistics education sketched above may seem to have self-evident rationales. Yet, I argue that they are insufficient, and carry hidden shortcomings, given the desire for skill transfer that underlies curricular guidelines worldwide in mathematics and statistics education, that is, the desire that education will contribute to effective life as adults. If our goal is to prepare future citizens and workers, we must examine the real-world demands on adults as they engage with models, so that we know what to prepare our graduates for—and design instruction from that basis, not just based on an internal lens as described earlier.

The issue of skill transfer has been recognized and researched over several decades in the fields of cognitive psychology and workplace training. From a cognitive psychology perspective, Perkins & Salomon (1992) argue that transfer includes "near transfer" (to closely related contexts or performances) as well as "far transfer" (to quite different contexts or performances); They further emphasize that research shows that often transfer—especially far transfer—does *not* occur.

The challenges of far transfer should bother scholars and educational systems (i.e., both schools and academic departments) interested in enhancing their graduates' understanding of models and modeling in the real world. The transfer literature argues that the tasks used during instruction should embody key elements that characterize the real-world tasks onto which transfer is expected. In contrast, the way models and data-based modeling are portrayed to the public in the media often differ from how it is experienced in the mathematics or statistics classroom. For example:

- Gal et al. (2022a) report on analyses of media materials conducted as part of the ProCivicStat project (see http://iase-web.org/islp/pcs) which suggest that Civic Statistics (including models) reported to the public are based in part on dynamic and multivariate data which are different than the one-shot (e.g., survey-based) statistics or simple datasets often used by teachers.
- Gal & Geiger (2022) present new findings regarding the cognitive demands of media items related to the COVID-19 (Corona) pandemic, based on a content analysis of 300 media items from four countries. They point to the fuzzy nature of some of the wording used in the media to describe models and the results of modeling; and also to "embedded criticality," that is, ways the media publishes critical views of data (e.g., by showing interviews with experts who disagree or by publishing different estimates of some value). As a result, readers need to understand the statistical basis of these disagreements and uncertainties and witness good and bad examples of how to be critical of data and models. These observations and others attest to the centrality of text comprehension and other capacities when interpreting information about models.
- The media makes extensive reference to statistical indicators, which are standard tools in official statistics and heavily reported by official statistics agencies (Gal, 2003) and often picked up by the media for further reporting to the public. Yet statistical indicators are seldom covered in introductory statistics courses (Pfeffermann, 2015). Such indicators are further discussed below.

The analyses above imply that teaching about statistical models that is based on an internal view of models and modeling cannot lay a sufficiently robust foundation for skill transfer. This chapter argues that as part of the search for ways to improve the pedagogy of teaching data-based models, there is a need to (also) adopt an external view of models and understanding of what is statistical modeling, and make sure learners are exposed in class to models and to the results of using modeling as they are used in the real world. An external view necessitates that we further examine the contexts within which data-based models arise in the real world, and their demand characteristics.

About real-world contexts for data-based models

As educators, we hope that what we have taught our learners will help them function effectively in multiple life contexts outside the classroom. However, as Gal (2023) argues, "context" is far from being a simple notion, because context is not automatically present in the classroom or lecture hall; educators need to bring it in. Going beyond the familiar adage that data are numbers in a context, it is essential to emphasize that understanding the context is mandatory when teaching statistical ideas, since it is the source for the "need to know" of different actors (e.g., governments, business organizations) which cause data to be collected in the first place and then analyzed. Further, the context informs the decisions about the methods (including models!) used to analyze the data, and without the context we cannot know what to analyze or how to interpret any emerging results, imbue them with meaning, or connect them to societal and

policy issues, which are the heart of Civic Statistics (Gal et al., 2022a).

However, when thinking about teaching that aims to enhance understanding of real-life statistical models, there is a need to focus on contexts that may either be familiar to and motivating for students, or of value in terms for preparing learners for their life roles as engaged citizens, smart consumers, and productive workers. In short, we need to make wise choices about which contexts to use. Here I sketch in broad strokes three separate rich functional contexts that may be useful in this regard. (Note: each has nuances that are not discussed due to space constraints):

A. *Models in media interpretation contexts.* People have to understand models and the results of modeling when reading or watching the news, broadly viewed, including newspapers and print media, websites of news organizations, posts on Facebook and X (Twitter), blogs, etc. Such channels routinely communicate statistical and mathematical products (StaMPs; see Gal & Geiger, 2022) that refer to models and the results of modeling, among other things. Model-related information is generated and shared by many actors: official statistics producers, public agencies, or researchers whose results are of interest to the media, such as regarding the progress of COVID-19, global warming, crime, health, equity, and other topics subsumed under Civic Statistics (Gal et al., 2022a).
B. *Models in service consumption contexts.* People (including young adults) encounter the results of models-in-use in diverse service contexts, when they act as customers of both commercial and public services. Examples are when people engage with an online shopping website or a social networks, where they encounter advertisements affected by algorithm-based technologies powered by various types of statistical models, or when they wait in telephone queues (e.g., when calling a call center). Service recipients normally have no access to the underlying data and system logic, and may not even be aware that models and modeling are operating "under the hood," but are heavily affected by the decisions and actions informed by such models.
C. *Models in workplace or employment contexts.* Employees in entry-level jobs or line managers (i.e., what school graduates may reach in the first few years) encounter results of models and statistical modeling in many work-related situations. Examples are when workers and managers are presented with and have to address KPIs (key performance indicators or metrics) in service centers (e.g., average wait time), in marketing (e.g., sales per hour, customer value, customer churn), or in operations (e.g., safety metrics). In addition, workers and managers may face many kinds of statistical predictions, such as regarding anticipated sales or production levels within a given number of months, breakdown or product failure forecasts, and the like. Such models are central in many organizations, since they are used to monitor performance and productivity of workers or departments, or inform e-recruitment and worker selection processes (Smythe et al., 2021).

These three contexts have some overlap, in part because the media reaches out and covers a wide range of topics, including those related to services and labor market issues. For example, the media may publish an article about the validity or fairness of employment acceptance decisions or discuss discriminatory effects on some social groups due to "algorithmic bias" (Barocas & Selbst, 2016) when human resources departments or employment agencies use algorithmic models to screen CVs, LinkedIn profiles, or online applications.

Characteristics of statistical models in the real world

An overview and pedagogical focus

Models and the results of modeling are present, in explicit or implicit ways, in the three central life contexts listed above (media interpretation, service consumption, workplace environments), yet these contexts seldom come into view in professional literature and teaching resources in statistics education on the teaching of models, even when educators claim to be concerned about skill transfer (Son et al., 2021). As noted earlier, extant pedagogies for teaching about *data-based* models and modeling focus mainly on analyzing raw data or on the mastery of computational routines related to statistical modeling (Zieffler et al, 2021).

Indeed, the majority of the chapters in this book, which are based in papers presented at the Minerva School 2022, demonstrate that extant approaches to teaching revolve around having students analyze raw data, whether "real" data in the sense of being authentic data, or cleaned or pre-fabricated datasets created to help the teaching/learning process. Certainly, engaging students with the statistical modeling cycle has its own logic and educational benefits, yet it does not eliminate the need to continue and be concerned about skill transfer as discussed above.

A core question that is therefore raised here asks: *What should students know about the nature of the statistical models used in the key contexts listed earlier, with which they will have to engage as adults, i.e., media interpretation, service consumption, and the workplace?*

Of course, this is a very broad question, whose full treatment goes beyond the scope of a single chapter. As a starting step, and to help plan effective pedagogical strategies in this regard, it is necessary to be selective, and to identify and focus on a few basic but high-level ideas that can fit multiple instructional settings, and serve to prepare citizens of the future. We must take into account the existence of time and space constraints in a statistics classroom, which usually aspires to cover many topics and is already packed; hence adding new elements that teachers need to introduce must be planned with overall balance in mind.

To provide a deeper focus for instruction that can increase the chance for skill transfer to real-life contexts as sketched above, I believe that it is necessary to refer to a few principled components of real world models. The ideas below reflect insights that have grown out of my cumulative work and prior and ongoing analyses of the demands of real world statistical and mathematical messages. Examples include analyses of the characteristics of products of official statistics agencies (e.g., Gal, 2003), the cognitive demands that serve as building blocks of "official statistics literacy" (Gal & Ograjenšek, 2017), the facets and tools needed to understand diverse types of Civic Statistics (Gal et al., 2022a, 2022b); and the nature of statistical and mathematical products communicated to the public regarding the COVID-19 pandemic (e.g., Gal & Geiger, 2022), and other efforts that have encompassed statistical models.

Three basic uses of statistical models

A key question being asked here is, what are the key *families* of models that citizens and workers are actually exposed to? The notion of "family" relates to the *purpose* that the model fulfills, to *why* we need it, not to the underlying method of computation. Of course, this issue has received ample discussions within the statistical community, such as regarding predictive models, inferential models, and other kinds of models (Gould, 2024, **in this volume on page 81**). Statistics textbooks mention many types of models when discussing specific statistical methods or techniques. However, the plethora of technical terminologies in this regard can be daunting for educators, and teaching time for seemingly new topics is limited. Educators have to carefully choose what to focus on in class, and also work in ways that take into account students' negative attitudes about the difficulty of statistics or a sentiment that it is irrelevant to their lives (Schau, 1995).

Thus, this chapter is based on the operating assumption that there is a need to *simplify* the terrain with which teachers have to deal with, and offer a *simple* view of models in terms of their usage or purposes, one that students can (more) easily relate to and feel is accessible.

For simplicity, and to reflect the types of models that citizens and workers arguably encounter in diverse contexts based on the background work described earlier, below I sketch three basic[1] *families* of models which often appear in the media or in organizational work contexts, and which are proposed as a focus for basic instruction on real-world statistical models.

A. *Aggregation models — used as tools to describe the status of key social or organizational phenomena.* This is arguably the most prevalent use of models. Think of any statistical indicator used by official statistics agencies (e.g., infant mortality, Gini coefficient, high-school graduation rate, R-rate or positivity rate for COVID) or of workplace or business metrics (e.g., average waiting time for incoming calls; worker productivity). There are hundreds of these indicators, each based on a combination of some raw variables, using a set of rules for computation that yield a single value — a percentage, ratio, or number on an arbitrary scale that can vary up or down. Each indicator or metric is a bona fide statistical model because it simplifies a complex reality; it reflects the status of a key target phenomenon of interest to policy makers or managers (and often to the general public); and it is statistical because it uses data, and the data varies (i.e., the indicator is never fixed, but can and should be recalculated time and again, as data changes or varies).
B. *Prediction models — used as to anticipate the future status of key social or organizational phenomena.* This is a traditional and well-known aspect of the use of statistical models. Researchers in diverse fields (e.g., economics, medicine, science), organizational analysts, and official statistics producers use correlational procedures and a range of regression and related multivariate models. The aim of such procedures, in its simple form, is to examine relationships between variables, determine to what extent certain target phenomena can be predicted (i.e., modeled) by selected predictor variables, and determine the strength and shape of that relationship. For example, in commercial service contexts, which were noted earlier, models may be used, among other things, to predict customer purchases (who buys more?), customer attrition (who will stop using our services?), customer behavior while waiting in queues, or levels of supply and demand. Overall, prediction models provide information and predictions that can aid decisions and policy-setting in critical areas.

1 There are of course additional types of models that serve other purposes and may be employed in more specialized circumstances, such as models for classification or clustering, for optimization, etc. These are intentionally excluded in this chapter, since the goal is to focus on a few selected types of models at different levels of simplification, which can fit diverse types of learners and teachers, and teaching contexts.

C. *Simulation models — used to forecast (related to prediction, but different), to generate scenarios, and to estimate risk levels.* Simulation models are usually more complex than regression models; they may involve many underlying *assumptions*; they may not necessarily be based on real data but on "realistic" data; the data may be complex (e.g., global warming simulations use historical records, dozens of variables); analytic methods may be very diverse, including data-mining or other complex statistical techniques. Further, the data may be manipulated intentionally (i.e., researchers run different scenarios) to test how different conditions affect the results of "what-if" situations. (e.g., "if all people keep social distancing of at least 2 meters as regulations require, and wear face masks, then what are the chances of infection, compared to…").

It can be claimed that these three families of models (in terms of their function) go up in their sophistication, i.e., from aggregation which is seemingly the simplest, to prediction models and then to simulation models which are the more demanding. However, this is somewhat deceiving, because aggregation models can (and do) sometimes serve the same purpose as prediction models! by tracking the rise or fall of indicators over time, we can predict where things may be heading. Further, aggregation models are not trivial at all. Aggregation models deal with topics of much importance in civic, economic, and organizational contexts, hence could be of much interest to learners and teachers alike. Since they are simpler than the other types of models, aggregation models may be prime starting points for planning future pedagogies and starting points to teaching about real-world models.

The three frameworks that underlie all models

The three families of models discussed above, i.e., aggregation models, prediction models, and simulation models, all share three underlying building blocks or "frameworks." These frameworks are central for understanding any type of model, and thinking critically about models and results of modeling reported in the media and in workplace contexts:

A. *The conceptual framework.* This framework describes and requires conceptual decisions about the nature (or components or definition) of the target phenomenon, that is, the social, economic, organizational, environmental, or other topic of interest that has to be modeled (i.e., described, predicted or simulated). Then, the conceptual framework also describes (more decisions!) *assumptions* about the variables that matter and have to be included in the model, as well as why they should be included, and measured in the way planned. This framework also involves decisions on whether data will be needed on *other* variables, which will not be included in the model, but serve as correlates, otherwise the model cannot be linked to other variables, and any related results cannot be properly interpreted. For instance, when the Ministry of Education plans to survey "violence in schools" (or students' attitudes towards mathematics, and so forth), the first step is to define (conceptualize) the target phenomenon. For example: What is "violence in school"? Does this include physical violence? Digital violence? Verbal abuse or bullying? Any other type of violence? Only after decisions are made in this regard, which are conceptual or qualitative decisions, not statistical decisions (Ograjenšek & Gal, 2016), can a measurement methodology be planned.

B. *The measurement framework.* This framework involves decisions on *how* to collect the needed data, and *why*. This covers the topics normally included in the methodology of a study, including data sources and methods, measurement instruments, etc. Most but not necessarily all of these topics are covered in a standard introductory statistics course or textbook, hence no reference is provided here. But note that the "data" is a broad term, which may encompass "objective" sources (e.g., administrative records) or survey-based "subjective" data, or even "big data." as well as textual data that can be categorized and quantified (as in texts of customer queries or complaints that is recorded on a company website). As an example, think of the difference between two well-known economic indicators: "consumer spending" vs. "consumer confidence," which are used to model and predict key economic trends and inflation. One of these is seemingly objective, the other one subjective, hence each requires a very different measurement methodology.

C. *The computational and reporting framework.* This framework involves two related but separate aspects: how to combine/integrate the different elements (variables) in the model, and how to report them to target audiences in a way that enables to derive meaning and insights from the results of the modeling effort. For example, the result of a model related to income may be measured and computed on a continuous scale (e.g., in terms of total net income, or gross income before taxes), but then related to "poverty" via a dichotomy of what percentage of the population falls above or below a "poverty line" (Note: the poverty line itself is a separate theoretical or simulation-based model of a critical social phenomenon). Many examples for computational and reporting frameworks can be found in press releases and reports by official statistics agencies (see Gal, 2003) which are then picked up by the media and reported to the public.

The construction of any model requires all the three frameworks described above. Hence, all of them must be addressed in instruction regarding data-based models

used in real-world contexts. However, the conceptual framework is the most critical! As Ograjenšek and Gal (2016) argue, any data collection effort has a qualitative core, shaped by the actors that initiate and decide what needs to be modeled, and why. This, in turn, determines the methods deemed acceptable and useful for reaching these goals. For example, in developing a prediction model, the most important decision is understanding what is the nature of the phenomena to be predicted, and why we want to predict it, and then, determine what variables to include in the model (or exclude). These are essential, so that we are able to connect the results to the social meaning and implications, a point emphasized by the ProCivicStat project (Gal et al., 2022a). The adequacy of the conceptual framework is also an important basis for interpreting the outcomes of the model and evaluating its credibility, well beyond "statistical" aspects included in the other frameworks. Such issues regarding the three frameworks listed above will be illustrated at the workshop in reference to "gender pay gap" and other key topics in the news media.

Discussion: implementation, critical thinking, research directions

This chapter has argued that a discussion about developing relevant pedagogies for understanding models and the results of modeling must adopt an external lens and consider the characteristics of the actual information about and characteristics of models and modeling with which citizens and workers have to engage out there in the real world. Good statistical thinking, and effective statistical literacy habits, which are among the key goals of statistics (and mathematics) education, can be developed not just by having students analyze raw data. The chapter innovates by sketching three types of models and three specific conceptual frameworks that, taken together, can define a stand-alone learning outcome, and can serve as a basis for designing curricula and instructional sequences that fit diverse teaching contexts.

The Minerva School 2022 was instrumental for me in developing and validating my core ideas, by providing access to papers and presentations about current efforts to conceptualize and teach about data-related models in a wide range of teaching/learning environments, within mathematics, statistics, and science education contexts. Through reflective discussions in small and large groups, the School's work format made it possible for me to compare my own views against those of others, and realize that the vast majority of current approaches revolve around having students analyze raw data (whether authentic or cleaned datasets). In contrast, the ideas sketched in this chapter point to the need to engage students with meaningful and authentic *texts* and *contexts* (Gal, 2023), with an emphasis on examples for models reported in the media and in service and workplace contexts. Specifically, the chapter argues that educators should focus on introducing learners to the existence of aggregation models, prediction models, and simulation models, and that they learn that underlying each of these types of models are a conceptual framework, a methodological framework, and a computational and reporting framework. These ideas go well beyond extant conceptualizations of modeling and its connections to citizenship roles (Maass et al., 2023).

The ideas presented in this chapter have to be viewed with caution, given that teaching at the high-school (or upper middle school levels) may involve quite different constraints and realities compared to university-level introductory statistics courses. Paradoxically, the school environment may offer teachers relatively more flexibility, since instruction on model-related issues can be part of instructional sequence that stretch across multiple school years, and connect with information about models and modeling presented in multiple school subjects. It may be possible to start with seemingly simple descriptive or aggregation models that involve only two variables, where the computation of the model itself may be simple and accessible to most students, and explain how tracking them over time may enable predictions into the future, an approach that can introduce students to the idea of prediction without using any complicated computations. In contrast, at the college level, adding instruction on real-world models to an already crowded introductory statistics class may be a challenge, time-wise, meaning that instructors may be forced to select only one or two carefully selected examples. For example, it may be possible to superimpose the conceptual frameworks introduced in this chapter (e.g., a conceptual framework, a methodological framework, and a computational and reporting framework) on existing content that pertains to the use of measures of central tendency (when learning about averages) or regression, and connect them to the use of real-world models for aggregation and prediction taken from media and official statistics sources.

Regardless of the teaching context and students' age and mathematical background, the ideas sketched in this paper can equip students to become statistically literate and *critical consumers* of data-based statements (Gal, 2002; GAISE, 2016). This involves, among other things, understanding and critical interpretation of messages about models and model-based results in a range of realistic contexts. There is room to further develop this area, which is also addressed by Büscher (2024, **in this volume on page 49**), for instance by more specificity regarding relevant "worry questions" (Gal, 2002; Wild and Pfannkuch, 1999) that pertain to all the three types of models and three frameworks underlying all models.

In sum, we must help all citizens develop their ability to evaluate models critically regarding these ideas and more:

- the conceptual framework that underlies the model,
- the adequacy of the assumptions made about the variables and the data (e.g., about linearity of relationships),
- the quality of the data used and its completeness,
- biases caused by the reporting framework, and related issues (Bailey and McCulloch, 2023).

Looking ahead, we still have to rethink how the frameworks described in this chapter can fit into and enrich extant curricular sequences, so that teaching about models can enable *skill transfer* and prepare learners for life roles as citizens and workers. Such curriculum design decisions will differ for high-school and for college-level introductory statistics courses, since each involves a different operational environment, yet have to expose students to meaningful contexts, texts, and questions about real-world statistical models (Gal, 2023). Educators should seek ways to merge traditional instruction that exposes students to the statistical modeling cycle and to existing computational methods (e.g., linear regression) with the principled ideas sketched in this chapter. However, further research in this regard is needed, perhaps involving a "research design" approach, in order to determine how to weave a regular teaching sequence in statistics with one that emphasizes conceptual understanding of real world models. In addition, the ideas introduced in this chapter can serve learners as thinking tools that help *critical evaluation* of information about models across a range of real-world contexts. Yet, there is a need to further study how learners grow and develop their ability to think critically about model-related information taken from the media or from reports by statistics producers, and their level of critical stance (Gal, 2002), i.e., preparation for and comfort with taking an actively critical role regarding model-based messages.

References

Bailey, N. G., & McCulloch, A. W. (2023). Describing critical statistical literacy habits of mind. *The Journal of Mathematical Behavior*, 70, 101063. https://doi.org/10.1016/j.jmathb.2023.101063

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671–732.

Blum, W., & Borromeo Ferri, R. (2009). Mathematical modelling: Can it be taught and learnt?. *Journal of Mathematical Modelling and Application, 1*(1), 45–58.

Büscher, C. (2024). **"Design principles for developing statistical literacy by integrating data, models, and context in a digital learning environment" in this volume on page 49.**

Cevikbas, M., Kaiser, G., & Schukajlow, S. (2022). A systematic literature review of the current discussion on mathematical modelling competencies: State-of-the-art developments in conceptualizing, measuring, and fostering. *Educational Studies in Mathematics, 109*, 205–236.

Doerr, H. M., & English, L. D. (2003). A modeling perspective on students' mathematical reasoning about data. *Journal for research in mathematics education, 34*(2), 110–136. https://doi.org/10.2307/30034902

GAISE (2016). *Guidelines for Assessment and Instruction in Statistics Education: College report.* American Statistical Association. Online: https://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx

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

Gal, I. (2003). Expanding conceptions of statistical literacy: An analysis of products from statistics agencies. *Statistics Education Research Journal. 2*(1), 3–22. www.stat.auckland.ac.nz/serj

Gal, I. (2023). Critical understanding of Civic Statistics: Engaging with important contexts, texts, and questions. In J. Ridgway (Ed.), *Statistics for empowerment and social engagement – teaching Civic Statistics to develop informed citizens (Chapter 13)*. Springer.

Gal, I., Grotlüschen, A., Tout, D., & Kaiser, G. (2020). Numeracy, adult education, and vulnerable adults: a critical view of a neglected field. *ZDM-Mathematics Education, 52*(3), 377–394.

Gal, I., & Geiger, V. (2022). Welcome to the era of vague news: a study of the demands of statistical and mathematical products in the COVID-19 pandemic media. *Educational Studies in Mathematics*, 1–24. https://doi.org/10.1007/s10649-022-10151-7

Gal, I., & Ograjenšek, I. (2017). Official statistics and statistics education: bridging the gap. *Journal of Official Statistics, 33*(1), 79–100.

Gal, I., Nicholson, J., & Ridgway, J. (2022a). A conceptual framework for Civic Statistics and its educational applications. In J. Ridgway (Ed.), *Statistics for empowerment and social engagement: Teaching Civic Statistics to develop informed citizens* (Chapter 3). Springer.

Gal, I., Ridgway, J., Nicholson, J., & Engel, J. (2022b). Implementing Civic Statistics – An Agenda for Action. In J. Ridgway (Ed.), *Statistics for empowerment and social engagement: Teaching Civic Statistics to develop informed citizens* (Chapter 4). Springer.

GAISE (2016). *Guidelines for Assessment and Instruction in Statistics Education: College report.* American Statistical Association. Available from: http://www.amstat.org/ASA/Education/Undergraduate-Educators

Garfield, J. B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). Learning to reason about statistical models and modeling. *In J. B. Garfield and D. Ben-Zvi (Eds),. Developing students' statistical reasoning: Connecting research and teaching practice* (pp. 143–163). Springer.

Gould, R. (2024). **"Traditional statistical models in a sea of data: teaching introductory data science" in this volume on page 81.**

Gravemeijer, K., & Doorman, M. (1999). Context problems in realistic mathematics education: A calculus course as an example. *Educational Studies in Mathematics*, *39*(1), 111–129.

Lesh, R., & Lehrer, R. (2003). Models and modeling perspectives on the development of students and teachers. *Mathematical thinking and learning*, *5*(2–3), 109–129.

Maass, K., Zehetmeier, S., Weihberger, A., & Flößer, K. (2023). Analysing mathematical modelling tasks in light of citizenship education using the COVID-19 pandemic as a case study. *ZDM–Mathematics Education, 55,* 133–145.

Mallows, C. (1998). The zeroth problem. *The American Statistician*, *52*(1), 1–9.

Ograjenšek, I. & Gal, I. (2016). Enhancing statistics education by including qualitative research. *International Statistical Review, 84(*2), 165–178. DOI:10.1111/insr.12158

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. *International Encyclopaedia of Education*, *2*, 6452–6457.

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

Pfeffermann, D. (2015). Methodological Issues and Challenges in the Production of Official Statistics (24th Annual Morris Hansen Lecture). *Journal of Survey Statistics and Methodology, 3,* 425–483.

Schau, C., Stevens, J., Dauphinee, T. L., & Vecchio, A. D. (1995). The development and validation of the survey of attitudes toward statistics. *Educational and psychological measurement*, *55*(5), 868–875.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Smythe, S., Grotlüschen, A., & Buddeberg, K. (2021). The automated literacies of e-recruitment and online services. *Studies in the Education of Adults*, *53*(1), 4–22.

Son, J. Y., Blake, A. B., Fries, L., & Stigler, J. W. (2021). Modeling first: Applying learning science to the teaching of introductory statistics. *Journal of Statistics and Data Science Education*, *29*(1), 4–21. doi: 10.1080/10691898.2020.1844106

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248.

Zieffler, A., Justice, N., delMas, R., & Huberty, M. D. (2021). The use of algorithmic models to develop secondary teachers' understanding of the statistical modeling process. *Journal of Statistics and Data Science Education*, *29*(1), 131–147.

Some reflections on the role of data and models in a changing information ecosystem

Joachim Engel

Ludwigsburg University of Education, Germany
engel@ph-ludwigsburg.de

The information landscape is changing dramatically in the digital age due to the increasing availability of information through the internet and the widespread use of digital technologies. With the abundance of data and the ease of access to data analysis tools, individuals (students, citizens, …) need to be aware of the limitations and potential biases of the data, as well as the limitations and potential errors of the statistical models and methods being used. It has never been more important than today to be able to judge the credibility of information and its sources. To increase our students' resilience against misinformation, manipulation and outright fake news we need to design our teaching in a way that supports critical and well-founded reasoning with data and models.

The information ecosystem and its pollutants

Digital media and the availability of data of almost unlimited scale are radically changing our access to information. New sources of data are providing new kinds of evidence, provoking new kinds of questions, enabling new kinds of answers, and shaping the way evidence is used to inform decision making in private, professional, and public life. The access to a wider range of information from a variety of sources, an increased efficiency, and enhanced connectivity facilitating people to cooperate regardless of geographic location promises to improve the quality of life. However, the overload of free and unchecked information also presents specific challenges to democratic and open societies. The ease of creating and sharing information has led to an increase in the spread of false or misleading information. The information ecosystem is a helpful metaphor in referring to the complex network of actors, technologies, and processes involved in the creation, distribution, consumption, and management of information. It encompasses a wide range of sources, from traditional media outlets and government institutions to social media platforms and individual users. Ridgway and Ridgway (2022) give a detailed account of the range of actors in the information ecosystem and the role they play, their access to evidence, and how they influence consumers with their messages. The information ecosystem is changing rapidly due to advances in technology and changes in the way people consume and interact with information.

Some elements in the information ecosystem hold great promise to create new knowledge and to help with making better decisions to improve quality of life, e.g., in medicine, education, or the economy. However, as in any ecosystem, there is an alarming level of pollution—toxic elements—that must be kept within bounds for any ecosystem to survive. In the information ecosystem, polluters are entities that spread false, misleading, or biased information, often with the intention of manipulating public opinion or achieving a specific outcome. Disinformation, fake news, alternative facts, and conspiracy theories are on the rise. While acting under freedom-of-speech laws, these polluters can have a significant impact and are a serious threat in democratic societies. For democracy to work, citizens must have a critical understanding of empirical evidence on important issues of social and economic well-being and human rights. Sound evidence-based decision making in both private and public life requires quantitative reasoning skills and (equally important) a positive attitude toward engaging with data. Implementing difficult decisions on controversial societal issues (such as migration, climate change, pandemics) or forcing behaviors that have profound effects on people's lifestyles (e.g., policies to curb the spread of disease) depends significantly on the consent and support of citizens. The ProCivicStat Project (Ridgway 2022, Engel 2017) provides a detailed framework for analyzing the demands on citizens in democratic societies to understand statistics about society, including guidelines for instruction and concrete materials for teaching.

In an increasingly data-driven world, social, societal, and technological change requires new competencies. In addition to the obvious technical ICT skills and basic statistical knowledge, this includes the ability to evaluate the suitability and credibility of data-based arguments and to reflect on the societal impact of technological solutions in an ethical way. It also includes the capacity to distinguish reliable data from fake news. This expansion of competencies affects not only the professional world, but all of us. Innovation, social progress, and the well-being of our civil society require that people in science, business, politics, and society know how to evaluate and make sense of data to develop a sound understanding of our world and address pressing societal challenges with empirical insights and sound data-driven arguments. At the same time, Big Data, with its possibilities for surveillance, manipulation, and control, raises serious problems for democracy and freedom (see, e.g., Helbing et al. 2017). Algorithms that draw on data are used to profile members of society and make important decisions that disproportionately impact those with less privilege and resources (O Neill 2016).

There are two factors that determine survival in a polluted ecosystem: the extent of the pollution and the resilience of the species. It has never been more important than today to be able to judge the credibility of information and its sources and to understand the role of data in creating new knowledge. As Ridgway & Ridgway (2022) state, "students need to be aware of the web of creation and destruction that underpins knowledge building." Two key ingredients in creating new knowledge are data and models. As educators, we need to look for ways to develop students' resilience—in particular, to make them more resistant to polluting elements in the evidence ecosystem. Resilience against false information requires a critical stance towards data and an awareness about the role of data and models in creating new knowledge. Key elements are a critical appreciation of data, their source, reliability, and appropriateness to address the issue under consideration. Equally important is a reflective appreciation and understanding of the role of models underlying any conclusions that were drawn from the data. Before discussing how to include an appropriate critical stance on data and models in our teaching, we consider three cautionary examples, and then look at the processes involved in transforming data and evidence into knowledge and new wisdom.

Three examples for poor data and misleading models

The following examples highlight flaws caused by uncritical (or intentional mis-) use of data and models resulting in misleading and wrong conclusions. They are intended to illustrate the importance of being on guard against misinterpretation.

1. **Poor data in predicting election results:** To predict the outcome of the 1936 American presidential election over 10 million people were asked by the Literary Digest magazine to mail in their preference between the two candidates, Alf Landon and Franklin D. Roosevelt. Based on this survey, a clear victory of 54% of the votes was predicted for the challenger Landon over Roosevelt, whom the poll gave only a 41% share of the vote. The actual election results turned out to be just the opposite, a clear victory for the incumbent, with 60% of the vote for Roosevelt and 37% for Landon.
2. **Misleading operationalization of variables;** The website of the Australian-based news portal news.com.au reported on Feb 11, 2017 that the Vatican, with 1.5 crimes per person per year, is the country with the highest crime rate in the world. Similar statistical measures identify Frankfurt as most dangerous city in Germany (albeit by far not quite as "dangerous" as the Vatican).
3. **Inappropriate model:** Following a recent study by Kuhbandner and Reizner (2023) about excess mortality and vaccination campaigns, the Online Portal "Die Achse des Guten" (translated: "the axis of the good") concluded that the only factor explaining the excess mortality was the vaccination campaign. The report24.news channel reported about this study under the headline "The number of deaths exploded in direct temporal correlation to the vaccination campaigns," and suggested a causal relationship in the sense of vaccinations being responsible for the excess death rates.

A detailed account of these examples is discussed in the appendix.

The hierarchy of knowledge

A useful model for understanding the process of transforming data into information and ultimately into knowledge and wisdom is the so-called DIKW hierarchy (see, e.g., Frické, 2018). It describes the progression of knowledge from raw data to wisdom (Figure 1). DIKW stands for Data, Information, Knowledge, and Wisdom.

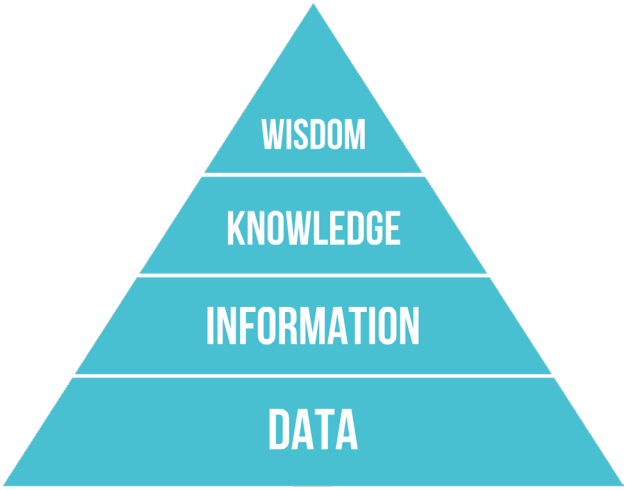


Figure 1: The knowledge pyramid or hierarchy of knowledge

Data are the foundational elements in the hierarchy of knowledge. They are viewed here as a representation of facts in a raw, unorganized fashion. Data values by themselves lack meaning and context. As Nate Silver's famous quote goes: "The numbers have no way to speak for themselves. We speak for them. We imbue them with meaning" (Silver, 2012).

The transition from data to information involves a process of organizing and processing raw data to make it meaningful and relevant. Through the identification of patterns, trends, and relationships, data become information. This process is enabled by abstracting from some irrelevant details in order to discern general structures and relationships in the data. It may include summarizing or reorganizing the data in an accessible and comprehensible way, creating a graphical representation of the data, or defining a new variable by operationalizing a concept. All of these actions require modeling activities. By interpreting and transforming information into a form that is useful for understanding, we create new knowledge.

The transition from information to knowledge involves a cognitive process where humans interpret and understand the organized information to form a coherent and meaningful representation of the phenomena under consideration. It requires individuals to process information in the context of their previous experience and integrate the information into existing knowledge. They connect the new information with what they already know and identify relationships and patterns between the new information and existing knowledge. This step of interpreting information to derive meaning and draw conclusions requires critical thinking and reasoning. It is influenced by our sense of finding and constructing meaning in the data-generated information and involves interpreting and understanding the relationship between different pieces of information. It also includes grasping underlying principles, concepts, and theories that govern the knowledge.

Finally, the transition from knowledge to wisdom represents a higher level of understanding and the application of knowledge in a thoughtful, ethical manner. It is reached by combining knowledge with experience, reflection, and judgment to arrive at a deep understanding of an issue or a situation. While knowledge involves understanding facts and information, wisdom goes beyond that to encompass judgement, discernment, and the ability to make sound decisions that promote well-being and positive outcomes. Wisdom also includes a critical reflection on one's own knowledge and experiences.

Notice, however, that the various steps in the hierarchy of wisdom do not follow a strictly linear order but are interdependent. It requires knowledge to draw valuable information from the data, i.e., which patterns to look for or which type of graph to draw as the most informative one, and it requires experience, judgment, and wisdom (the highest level in the hierarchy of wisdom) to support the interpretation of the information in a way that generates new knowledge.

The DIKW hierarchy is a model that highlights the importance of transforming data into useful information, and of combining knowledge with experience and judgment to arrive at a deeper understanding of a subject. By understanding the DIKW hierarchy, individuals can better manage the process of transforming data into knowledge and wisdom, and hence can make better decisions.

Notice the subjective element involved in each step of the knowledge pyramid.

Data–the raw material of statistics

Data—the empirical basis for evidence-informed decisions and knowledge creation—are certainly preferable to anecdotes, wishful thinking, superstition, prejudice, or ideology. Yet data themselves are neither facts nor truth. Some authors consider data as models of reality (Podworny & Frischemeier, **in this volume on page 15**). Data do not provide objective representations of the world. They might arise opportunistically, or as a result of conscious decisions someone made to research a particular topic. Data usually have been collected at costly expense, for a particular purpose and with a specifically

chosen research design. They measure manifest variables in a particular way. They are the basis for constructing latent variables based on some kind of model with a specific concept in mind. At a more complex level, one can ask why particular measures have been chosen, by whom, and for what purposes. Measurement is always linked to some theory of the phenomenon being studied. In the example of physical sciences, mass, length, and time were not chosen as measures because they are "obvious," but rather because when they are measured, precise predictions can be made and used about the physical world. The well-being of nations had been measured by the GDP per person until this monolithic measure was challenged by Amartya Sen; many people wish GDP were replaced by the more comprehensive Human Development Index.

Collecting data is not a leisure activity but is laborious, sometimes tedious work that usually requires a lot of effort and financial resources. It serves someone's interest, and it is legitimate to question whose interest this is. Why have these data been collected? The data collected implicitly tell a story. Whose story is this? And whose story is this not?

Critical or reflective questions about the methods used in surveys might include (but are not limited to):

- Are the measures (e.g., a questionnaire) well defined? Are the measures robust and appropriate for the purposes for which they are being used?
- Are metadata (i.e., detailed explanations of how variables were defined, sample characteristics etc.) available?
- Were the sampling procedures appropriate? Who is missing from the collected data? (e.g., measuring how citizens feel about a certain topic by analyzing social media streams fails to sample non-users).

Many studies in the social sciences are concerned with theories of causality; causality is associated with difficult philosophical challenges that go well beyond simple mantras such as "correlation does not imply causation." However, when data come from observational studies, surveys, or archive data, and not from experimental studies, a reliable identification of cause-and-effect relationships can be difficult to determine.

Beyond technical knowledge about processes of data generation, it is important that individuals are able to ask critical questions to assess the credibility and validity of any data, finding, or conclusion they encounter, both on technical and logical grounds—even data or reports from presumably credible sources such as official statistics agencies. It is important to examine, from a critical perspective, narratives and interpretations of data, and the conclusions drawn from them, for example:

- What is the quality of the evidence presented in a media article or a claim to support assertions about needed policy or actions (e.g., regarding recycling laws, wage equality, or vaccination)?
- How reasonable are the projections and how appropriate are the underlying statistical models and assumptions that have been applied to analyze data on key issues (e.g., on the progression of global warming or the rate of spread of infections such as the COVID-19 coronavirus pandemic)?
- When assertions are made about a correlation between variables (e.g., smoking and risk of death), are relationships assumed to be linear, and are they really so (or perhaps curvilinear)? More important, if causal processes or cause-and-effect relationships are assumed, are there plausible rival accounts, covariates, or unexplored intervening factors which could affect the findings?
- Are the conclusions consistent with other available evidence? When proposals are made for social policy, one can ask if the problem identification has been done adequately and whether relevant data have been used.

Fact-checking organizations are helpful in assessing the trustworthiness of data-related reports in the news (for a comprehensive list, see https://en.wikipedia.org/wiki/List_of_fact-checking_websites). The UK-based organization Full-Fact provides a toolkit to spot bad information (https://fullfact.org/toolkit/) and recommends asking questions such as

1. Where is it from? A trusted source is your safest option. If there's no source, search for one. If it doesn't look right, be careful.
2. What is missing? Get the whole story, not just the headline. Images and videos can be faked. Check what other people say.
3. How do you feel? People who make false news try to manipulate your feelings. If it looks too good to be true, it probably isn't true. Don't be the one who doesn't spot the joke.

Everyone needs to adopt a questioning attitude, and to know what questions to ask about the nature, limitations, or credibility of different data sources, statistical messages, and conclusions. But a critical stance when assessing evidence does not mean simply "blind" criticizing. Rather, criticism is about adopting the attitude of a fair skeptic who is ready to accept an account, but has to be convinced by evidence. In situations where data are presented in a misleading way, students should be encouraged to re-present them in more appropriate ways; in situations where data are dubious (or fabricated) students should be encouraged to find relevant data from authoritative sources.

Models–all wrong, but potentially useful

The core field of statistics is the application of models to represent situations of interest, e.g., to estimate the magnitude of a particular phenomenon or to forecast its evolution over time. Recent examples that have preoccupied the public discourse are the attempts to predict the progression of diseases during the COVID-19 pandemic or to forecast the pace of global warming or climate change. Such predictions are important to inform national policy decisions in this area. Modeling activities permeate the entire hierarchy of wisdom. Konold, Finzer, and Kreetong (2017) consider data as models (of reality). Moving upwards in the knowledge pyramid implies modeling at various stages. Information and new knowledge from data often involve building models to represent patterns and relationships in the data. Models help integrate information and knowledge from different sources and disciplines. They allow us to synthesize complex data and theories into a coherent framework.

Models reflect the perspective and interest of the model-building individual. The modeling process includes many subjective elements. For example, an economist and a sociologist might have quite different theories and methods for defining and studying poverty in society, and they may create different indicators to sum up different components that describe or predict poverty. Their models of the causes of, and remediation for, poverty might be quite different. Models do not represent an "objective independent reality" but only certain aspects while neglecting others, Therefore, following Büscher (**in this volume on page 49**), the reader of a statistical argument needs to interpret how the construction of the argument has been influenced by subjective perspectives of the contender.

The goal of statistical models is to extract insights and knowledge from the data to support decision-making, forecasting, and problem-solving. Traditional statistical models are used for deeper understanding to explain the phenomena observed while more recent models of machine learning focus on optimal prediction while ignoring the data generating process (Breiman 2001, Gould, **in this volume on page 81**). Models, by their nature, are not the real thing, but a simplification of the complexity and disorder that reality throws at us.

To simplify reality, models sacrifice details. Hence, discrepancies between the model and reality—the residuals—are not necessarily an indication of the model being inappropriate or useless. The analysis of the residuals provides information on whether the model can be held onto, i.e., whether it has proven itself, or whether the model is unsuitable. They are often a key to obtaining a deeper understanding of the phenomenon under investigation and perhaps to developing an improved model. For an appropriate model, the residuals should appear "reasonably irregular" (Tukey, 1977, p 549). Otherwise, the model can be improved iteratively by "adding" structure to the model in the residuals. This concept pervades all classical statistics, from univariate, bivariate, and multivariate data analysis, to data of all scale levels, independent (iid) data, or dependent (e.g., time series) data.

An example for inadequate modeling is presented by the Mackinac Center for Public Policy. "For most people, Coronavirus presents similar risks as car accidents", they claimed. In 2020, more people in Michigan died in a car crash than from COVID-19, they continued their reasoning, ignoring simple facts about infectious diseases. A virus spreads through human contact; therefore, all other things being equal—in particular, the absence of any intervention policy—the number of infected people will grow exponentially over time, unlike the number of traffic casualties.

Students need to acquire the ability to identify and understand the use of models, and to be able to challenge the fundamental assumptions made by any model. Overall, data and models are interconnected and synergistic components in the hierarchy of knowledge. Data provide the raw material from which information is derived, and models help organize, interpret, and extend that information into a deeper understanding of the world around us. Both data and models are essential tools for advancing our knowledge and making informed decisions in various fields of study and practice.

Conclusion

Citizens need to be empowered, and have skills in critiquing and interpreting evidence. Awareness of the role of data and models in knowledge generation plays a crucial role in the information ecosystem, as it helps individuals critically evaluate and interpret the information they encounter. Resilience requires critical thinking and rationality. Some key ways in which a critical appreciation of data and models can contribute to a more resilient information ecosystem include:

- Understanding the sources of data—to better equip learners to evaluate the credibility and reliability of data sources, and to identify sources that may be biased or misleading;
- Interpreting data—to enable individuals to interpret data and statistical analyses correctly, and to avoid common misconceptions and fallacies;
- Detecting misinformation—to help individuals detect false or misleading information that is presented using data or statistical analysis;

- Questioning the validity of model assumptions—to check if they are reasonable and if there are alternative assumptions;
- Validating the model—to test if based on new data or tough simulations the model leads to similar conclusions;
- Promoting transparency—to promote transparency in data analysis by advocating for the use of open data sources, clear methodology, and replicable results.

The evidence ecosystem will never be without some pollution. Keeping the level of pollution in the information ecosystem in check (while preserving democratic freedoms such as freedom of expression) is a challenge for society as a whole, not least for its legal and political system. Our task as educators is to enhance our students' ability to recognize and appreciate the broad context in which evidence emerges, and to strengthen their resilience to false information and misleading conclusions.

References

Binder, K., Krauss, S., & Krämer, W. (2019). Sonderbare Avocado-Vermehrung und kriminelles Frankfurt – Aktuelle statistische Fehler in den Medien unterrichtlich nutzen. *Stochastik in der Schule* 39 (2), 11–21

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199-231

Büscher, C. (2022). **"Design principles for developing statistical literacy by integrating data, models, and context in a digital learning environment" in this volume on page 49.**

Engel, J. (2016). Statistical Literacy for active Citizenship: A Call for Data Science Education *Statistics Education Research Journal*, 16(1), 44-49

Engel, J. (2017). *Promoting Understanding of Statistics about Society*. IASE Roundtable Conference: Berlin, Germany, 2016 Conference Proceedings https://iase-web.org/ Conference_Proceedings.php?p=2016_Promoting_Understanding_of_Statistics_about_Society

Frické, M.H. (2018). Data-Information-Knowledge-Wisdom (DIKW) Pyramid, Framework, Continuum. In: Schintler, L., McNeely, C. (eds) Encyclopedia of Big Data. Springer, Cham. https://doi.org/10.1007/978-3-319-32001-4_331-1

Gal, I., Nicholson, J., and Ridgway, J. (2023). A conceptual framework for Civic Statistics and its educational applications. In J. Ridgway, (Ed.), *Statistics for empowerment and social engagement: Teaching Civic Statistics to develop informed citizens*. Springer.

Gould, R. (2024). **"Traditional statistical models in a sea of data: teaching introductory data science" in this volume on page 81.**

Helbing, D., Frey, B., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R., & Zwitter, A. (2017). Digitale Demokratie oder Datendiktatur. In: C. Könneker (Ed.), *Unsere digitale Zukunft*. https://doi.org/10.1007/978-3-662-53836-4_1

Kish, L. (1965). *Survey sampling*. Wiley.

Konold, C., Finzer, W., & Kreetong, K. (2017). Modeling as a core component of structuring data. *Statistics Education Research Journal*, 16(2), 191-212.

Kuhbandner C, & Reitzner M (May 23, 2023) Estimation of Excess Mortality in Germany During 2020-2022. *Cureus* 15(5): e39371. https://doi.org/10.7759/cureus.39371

Lusinchi, D. (2012). "President" Landon and the 1936 Literary Digest Poll. *Social Science History* 36:1 https://doi.org/10.1215/01455532-1461650

O'Neill, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown.

Podworny, S., & Frischemeier, D. (2024). **"Young learners' perspectives on the concept of data as a model: what are data and what are they used for?" in this volume on page 15.**

Ridgway, J. (2022). *Statistics for empowerment and social engagement: Teaching Civic Statistics to develop informed citizens*. Springer.

Ridgway, J., & Ridgway, R. (2022). Civic Statistics in context: mapping the global evidence ecosystem. In J. Ridgway, (Ed.), *Statistics for empowerment and social engagement: Teaching Civic Statistics to develop informed citizens*. Springer.

Silver, N. (2012). *The Signal and the Noise*. The Penguin Press

Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison Wesley.

Appendix

Some more background information to the examples above

Poor data in predicting election results:

A historic example for using poor data refers to polls preceding the 1936 American presidential election (Lusinchi, D., 2012). To predict the election outcome, the magazine "The Literary Digest" sent out 10 million questionnaires to its subscribers. Based on the survey, a clear victory of 54% of the votes was predicted for the challenger Alf Landon over Franklin D. Roosevelt, whom the poll gave only a 41% share of the vote. The actual election results were a clear victory for the incumbent, with 60% of the vote for Roosevelt and 37% for Landon. Possible causes for this serious miscalculation were the so-called "selection bias" (*Who reads The Literary Digest? Are subscribers representative of all voters?*) and "non-response bias." Of the 10 million questionnaires, only 2.4 million were returned. Aren't the dissatisfied more likely to respond? A good quality survey must make every effort to keep the response rate as high as possible. Today, professional surveys to explore political behavior (elections, political opinions, party preferences) usually ask only about 1,000 to 1,500 eligible voters (Kish, 1965). Nevertheless, very precise results are usually achieved. Mathematically, it can be shown that with a simple random sample, as few as 1067 respondents are sufficient to determine with 95% certainty the true proportion within the population, with a margin of error of 3%. This is true if the true (but unknown) proportion is 50%. If this proportion is different from 50%, smaller samples are sufficient to achieve the same level of precision. In the Literary Digest survey, we succumbed to the misconception, which is also common among many students, that the most important thing is to have a lot of data. In contrast, it is better to have less good data than a lot of bad data. The random error due to small samples can be estimated and controlled, while systematic errors can hardly be corrected afterwards.

Misleading operationalization of variables

Several German newspapers reported that the most dangerous city in Germany is Frankfurt (see Binder et al., 2019). With 14,864 reported crimes per 100,000 inhabitants, this city led the German crime statistics in 2017. But is it really so dangerous? about 300,000 people commute to work in Frankfurt every day, and about 60 million passengers arrived at or departed from Frankfurt Airport in 2017. All the crimes committed by or suffered at the hands of these people are the responsibility of the city of Frankfurt. In Munich, by contrast, the airport belongs to the districts of Erding and Freising. For a meaningful comparison of crime across municipalities or countries, it would be better to relate the number of crimes to the number of potential victims and perpetrators rather than to the number of reported inhabitants.

The Vatican City State—apparently the most criminal country in the world—shows what absurd results crime statistics based on population figures can lead to. As "Radio Vatican" reports, there were a total of 640 civil and 226 criminal cases there in 2011—significantly more than one per Vatican citizen (492). But in 99 percent of the cases, not these, but one of the approximately 18 million visitors annually were involved as victims or perpetrators. Petty crime is greatly increased in the Vatican, just as it is in most of the world's tourist hotspots; the only difference is that the other tourist hotspots are not a state in their own right.

Wrong model connecting COVID vaccinations and excess mortality

A recent study by Kuhbandner and Reizner (2023) found higher excess mortality in the second and third years of the pandemic than in 2020 in Germany, with the increase correlating with the start of the vaccination. In contrast to the previous year, a high number of excess deaths was also observed in the months with a high number of first, second and third vaccinations. The temporal relationship between vaccination histories and excess deaths is particularly pronounced for the third vaccination. In September and October 2021, the initial small increase in the number of third vaccinations was accompanied by a comparatively small increase in excess deaths. In November and December 2021, the number of third vaccinations increased sharply, accompanied by a comparatively large increase in excess deaths. The report24.news channel reported about this study under the headline "The number of deaths exploded in direct temporal correlation to the vaccination campaigns," and suggested a causal relationship in the sense of vaccinations being responsible for the excess death rates. Another online Portal, "Die Achse des Guten" (translated "the axis of the good") explicitly concludes that the only factor explaining the excess mortality is the vaccination campaign.

The central problem of this analysis has been known in statistics for almost 100 years: the problem of so-called "Spurious Correlations" or nonsense correlations. This is based on the insight that when comparing two so-called non-stationary time series (i.e. time series with a trend), as Kuhbandner also did, high correlations are obtained even if there is no correlation between these time series.

Afterword: what we mean when we say "modeling"

Tim Erickson

Epistemoligical Engineering
eepsmedia@gmail.com

This afterword is a personal reflection about data modeling and about modeling in general, inspired by the work in this book. I'm going to focus on an issue that has been bothering me for a number of years: we use the words "model" and "modeling" for many different concepts, and I worry that this can be confusing for the people we most want to help: students and teachers. We will journey together through several varieties of modeling in order to demonstrate the variety of meanings, and then I will make a few observations and suggestions.

I love modeling.

In the introduction to this volume, following Mary Hesse (1962), we define a model as "a representation, an analogy, with a descriptive, explanatory, or predictive purpose." (page 7). We also assert that models frequently involve simplification, and that they are subject to review and revision.

For me personally, the most exciting part of that definition is simplification—*purposeful* simpification. A model airplane, made of plastic, is not the real airplane. The real plane is too complicated, heavy, and expensive. But my *purpose* is to have an affordable object with some salient attributes: the shape of the wings, the color and decoration—basically, the overall "look." I don't mind that it does not have ailerons or a working altimeter, or that it can't fly on its own, because this simplified plane fits my purposes.

If you had asked me for an example of *mathematical* modeling in 2010, I would immediately have picked a curve or a line on a scatter plot, a function to approximate the pattern in a set of bivariate data, like in Figure 1. The line *simplifies* the relationship, focusing on the signal at the expense of the noise. It smooths over any variability. We use that functional relationship to make predictions and develop insight into the phenomenon that produced the data. We can (and should) assess our models by comparing their results with reality, thinking carefully

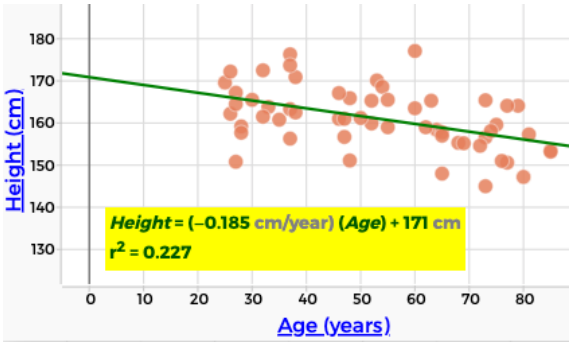


Figure 1: A least-squares linear model for the heights of 56 adult women as a function of age. Data from NHANES, display in CODAP.

about the importance of un-modeled factors, the conditions under which our models work reasonably well, and the possibility that the signal we see actually arose from chance alone.

As a description of the dataset, the line in Figure 1 captures the negative trend, smoothing over the variability. The line simplifies the data, giving us an elegant summary. Certainly this points-and-function activity is an example of *data modeling*.

Joachim Engel and I (Engel 2010; Erickson 2016) have independently produced collections of problems and activities focused on exactly this kind of modeling. The chapters in this book do not dwell on it, because data modeling is a lot more than functions on scatter plots, and the purpose of this book is partly to expand our understanding of what data modeling includes.

Now, if data modeling were the only kind of modeling, this book could be the definitive statement on the subject. But in fact, we use the words "model" and "modeling" in many ways, even within our limited scope as data and stochastics educators. As we continue our brief journey around the modeling landscape, I hope you will enjoy the variety we will see; but I hope you will also sense the "semantic peril" that I do.

Probability models

Let's start close to home. In the stochastics education world, we often talk about *probability* models. One web site (Yale, 1998) defines a probability model to be "a mathematical representation of a random phenomenon. It is defined by its sample space, events within the sample space, and probabilities associated with each event."

You can see how this definition leads to the familiar probability rules, using both theoretical and empirical approaches. But unless we squeeze that definition and twist it into shape, it's not "a representation, an analogy…" and so forth, but rather an intellectual paradigm for defining probability and making suitable calculations. There is nothing wrong with calling that a model—it's just something different.

For us, however, when we say "probability modeling," I think we usually mean something that *does* fit that earlier definition. It builds on sample spaces and all that, but it's more practical. Consider this tentative description:

> A probability model is a process, a set of connected stochastic events designed to produce results whose distribution and behavior mirror some other, possibly real-world phenomenon.

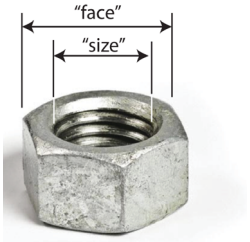
That is, we are making a model *of* something or *for* something. We use these models to calculate probabilities and to simulate data. Such a model simplifies a rich, uncertain process, and we would look at its results to see if they really work—for example, whether the simulated data make sense and resemble reality.

So if we had a problem about rolling dice and needed a probabiity, we could construct a probability model—either theoretical or empirical—to compute that value. Similarly, we talk about binomial models and Normal models as shortcut terms for certain distributions and procedures. We could talk about randomization procedures as using probability models, perhaps as a synthesis with data modeling. After all, when we use a bootstrap, say, we are using data within a stochastic process we have designed ourselves in order to produce a sampling distribution.

But there is also another, subtler "probability model." Suppose we flip a coin ten times and count the number of heads. We could use a binomial model with a probability of 0.5. But why use 0.5? In reality, there is no such thing as a fair coin. No coin has a P(heads) of exactly 0.500. So we simplify things and accept 0.5 as good enough. This observation suggests that *any time we assign a probability, we're modeling*. This is very much like the audacious suggestion in this book (Podworny & Frischemeier 2024, **page 16**) that data values themselves are models.

More kinds of modeling

Now let's think about *numerical* models. When we need to understand a complicated system—a system of differential equations, say, governing the motion of a spacecraft, or the relationships of forces in a truss, or the ebb and flow of temperature and humidity in the atmosphere—we often solve the underlying equations numerically because we don't know how to solve them analytically. The "numerical" part of the label is partly about the solution technique, and partly to signal that the answers and procedures are approximate; we are hoping the precision is good *enough*. They are models for that reason, and also because they result from choices we make about what to include in the system and at what precision: we decide what to model, simplifying reality.



Let's switch gears now and talk about what I'll call a *geometrical* model. Suppose we have data about some hexnuts, and we want to calculate the density of the metal. We have the mass, so we need the nut's volume. We have its dimensions, so we model its shape as a hexagonal prism with a circular hole. This is not *exactly* the shape of the nut, but rather a simplified shape: it doesn't account for easing of the edges or the threads in the hole. Like so many models, however, it captures the overall pattern, the relationship of the dimensions to the volume, so that we can make a good—but not perfect—calculation of the density. This kind of modeling shows up all the time in science and engineering. Rob Gould uses geometrical approximations like these when he talks about the volume of trees (**page 84**).

Even when we're not talking about data, however, we often use geometrical models to help with instruction or even our own understanding. Figure 2 shows a

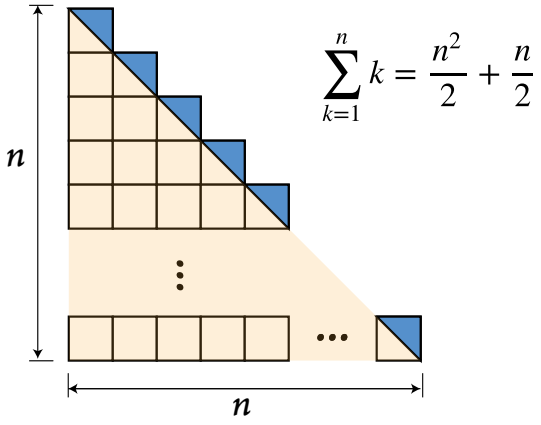


Figure 2: Area model to help explain a formula for triangular numbers. The large, light brown triangle is the $n^2/2$; the blue "teeth" are the $n/2$.

diagram—an *area model*—that helps explain a formula for triangle numbers. Note that this kind of modeling is not about simplification at all. There is no glossing, no approximation. The model includes *everything* from the context, yet it is still a model.

We use "pure math" models in stochastics as well, for example, if we use an area model to visualize conditional probabilities and Bayesian situations (e.g., Erickson 2017). Indeed, Karin Binder might have used such an area model (private communication), but the small sizes of some of the probabilities involved make her trees (**page 25**) or her net diagrams (**page 26**) more practical in her work.

We would probably all agree that both kinds of geometric modeling require important skills that we would like our students to have. (And that our students do not yet have these skills...) We'd also probably agree that this modeling is different in character from most of the modeling we have written about in this book. It "smells different."

Let's go farther afield. Tom Bielik (**page 33**) showed us a tool—Sage Modeler—for doing *systems* modeling. Internally, it's an example of a numerical model, but I want to shed light on it because we also use the term *systems model* to mean something more like the *diagram*—the boxes and arrows—and what they represent. Can a diagram be a model, even if it doesn't have quantitative information? Certainly! It's a representation of a complex process, made understandable and useful—in this case, through simplification. Implementing it in Sage Modeler makes it quantitative, and comparing those results to reality lets us evaluate and update the model.

Stepping even farther away from data modeling, let's look at another nodes-and-arrows model for a process: the PPDAC cycle (Figure 3).

Not only is this not quantitative, it's not even designed to help us analyze data from a real-world phenomenon. Instead, it's kind of a meta-model, helping us recognize the thinking and processes we go through when doing such an investigation. Is it actually a model according to the definition in the introduction? (**page 7**) Of course it is. It simplifies the investigatory journey, glossing over the dead ends, the backtracking, the flights of inspiration, the seeing what we will need three steps from now. A real path among the nodes in the diagram looks much more chaotic—but this simplified, circular diagram shows us the overall flow of the work and thinking. It's a kind of best-fit, optimized route, ignoring the variability, serendipity, and surprise.

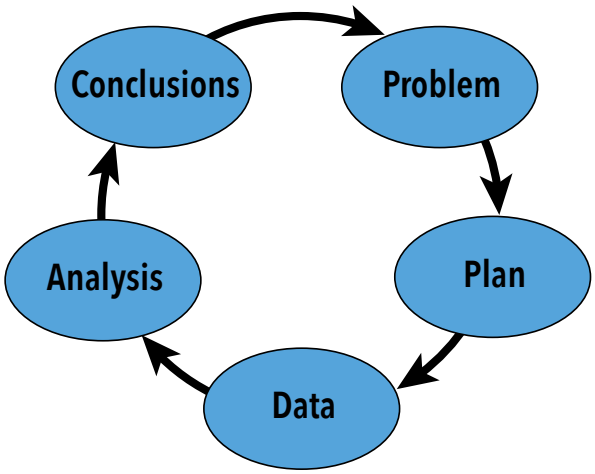


Figure 3: The bones of the PPDAC cycle, from Wild and Pfannkuch, 1999.

Reflection

I have mentioned six kinds of modeling: data modeling, probability modeling, numerical modeling, geometrical modeling, systems modeling, and finally a kind of conceptual modeling. There are more.

Whatever kind of modeling you're using, it's useful. It's exciting. And it connects dry math techniques to the practical real world. It is not a surprise that education leaders and organizations are calling for more modeling in all sorts of quantitative fields.

And yet the terms *model* and *modeling* mean many different, important things. In our roles as teacher educators and educational advocates, we want to promote modeling—but I guarantee that teachers will be justifiably confused about what we mean by the term. Imagine, for example, a workshop where we talk about trees as data models (e.g., Gould, **page 86**), and in the next session talk about the PPDAC model for investigations (Hagenkötter et al, **page 43**). Or where we explain how to use a net diagram as a model for conditional probability (Binder, **page 26**), but also try to explain that any measurement, recorded as data, is also a model (Podworny & Frischemeier, **page 16**). And it should not be lost on us that Gafny and Ben-Zvi (**page 69**) created what looks like a model of...data modeling pedagogy. Instructing teachers explicitly in the various distinctions and definitions that we discuss among ourselves may not help matters.

I don't have clear or comprehensive advice about this, but here is a suggestion: We ought to develop a variety of strategies for talking about modeling with regular people (such as teachers): people who have not had week-long discussions on the topic with an international group of thoughtful experts.

Here are some suggestions for things we should be thinking about:

- Teachers will be able to internalize the broad picture of what we often mean by a model: a representation of some system, subject to revision, and often (but not always) involving simplification. Teachers can understand that:
  - We use models because it's often easier to solve a problem in the model than in in the context directly.
  - Modeling happens when we recast a context or a problem in a mathematical way, that is, when we *mathematize*. In a student investigation, this often begins when they get data, and may undergo several layers of modeling as we do deeper and deeper analysis.
  - Later in the process, students must *de-model* when they reinterpret the results of an analysis and make sense of it in the original context.
  - Being willing to revise a model in response to that interpretation is an essential part of modeling.
- There are parts of a good investigation that are *not* modeling, and we should agree on what they could be. Otherwise, we might start to see everything as modeling, and that's not useful. One possible modeling-free zone happens between the modeling and de-modeling, when any computational analysis takes place. Note that a traditional curriculum spends much of its time there.
- We need to be patient and flexible. Modeling in schools will take many forms, at different levels of engagement. They are all legitimate. Some classrooms will only be able to incorporate a single activity with modeling; others whole units; others entire curricula where modeling takes center stage.
- The multidimensional framework from our introduction (**in this volume on page 10**) might be a bit much for teachers to digest (and keep separate in their minds from other flavors of modeling). But *we* need it, and so do other curriculum developers. It can become a powerful tool for thinking about data modeling in education. When we develop activities or think about learning trajectories, we should refer to it, to see what ideas we are incorporating, what we are setting aside, and what we might inadvertently be missing.

Let me close this personal reflection with a personal note of gratitude. I attended the Minerva School in 2022. After the event, I was asked to do some (mostly) gentle editing and to lay out this document. It has been an honor and a pleasure to do so, and I could not have wished for a more thoughtful, generous, and responsive group of authors, collaborators, and trusted friends.

References

Engel, J. (2018). *Anwendungsorientierte Mathematik. Von Daten zur Funktion*. Heidelberg: Springer-Verlag, 2nd Edition

Erickson, Tim. (2017). "Beginning Bayes." *Teaching Statistics*. (39) 1, 30–35. Winner of the Peter Holmes prize, 2017. Also appears as "Ein Zugang zu Bayes." *Stochastik in der Schule*. (3) 37, 30–34. Trans: Hans-Dieter Sill.

Erickson, T. (2016). *The Model Shop, Volume 1: Functions from geometry*. Oakland, CA: eeps media. Excerpted and translated as Erickson, T. 2019. *Mercado de la Modelación*. Oakland, CA: eeps media.

Erickson, T. (2014). Nuts, Fish, Babies, and Nuts. In Sprösser, et al., eds, *Daten, Zufall, und der Rest der Welt*. Wiesbaden: Springer Spektrum.

Hesse, M. B. (1962). *Forces and fields: The concept of action at a distance in the history of physics*. Mineola, NY: Dover.

Wild, C.J. & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*. 67, 3. 223–265.

Yale Univerity, statistics 101. (1997). http://www.stat.yale.edu/Courses/1997-98/101/probint.htm.

Postscript: Language Note

As you read in the introduction, the Minerva School is a joint Israeli-German project. What language should it operate in? English. Of course. As a native speaker, I benefit from the cultural hegemony my language enjoys; and I marvel at the depth and quality of the English of my colleagues when, of course, I would be baffled if the meeting were held in German or Hebrew.

This was brought brutally to my attention when I was editing the chapter by Susanne Podworny and Daniel Frischemeier. On **page 16**, Susanne was saying that the term *data model* would be a *pleonasm*—and I had never heard the word! I had to look it up. Where did she learn it? She explained that the word was in an advanced English course and she thought it the *mot juste*.

D'accord. But since this book is about data, I did a survey. I wrote to a few dozen of my most literate friends across the former British Empire, and *none* admitted to having a clue. One wag, after looking it up, wrote that he would rather have a pleonasm than a neoplasm.

For the record: a *pleonasm* is a word or phrase which is redundant or repetitive, though it can also be used for emphasis: e.g., "he was consumed by the *burning fire* of love"; or, perhaps, "*data model*." A *neoplasm* is a tumor or other abnormal and excessive growth of tissue.