# Taylor and Francis Book Proposal for Data Science in Education

By Emily Bovee, Ryan Estrellado, Jesse Mostipak, Joshua Rosenberg, and Isabella Velásquez

## 1. Statement of Aims

School districts, government agencies, and education businesses are generating data at a dizzying pace and serving it to teachers, administrators, and education consultants in a mind-boggling variety of formats. Educators and educational data practitioners who want to use data to improve the educational outcomes of students often have a clear idea of the questions they want to ask of their data, but they are left to analyze data as it is presented to them, often times using high-cost proprietary systems. Educational data rarely comes in a “ready-to-analyze” format, so educators and educational data practitioners who are eager to leverage data to promote student success often feel very little connection between the analytic questions they have and the numbers on their laptop. But some educational data practitioners are adopting the tools of data science, including open source projects like R, to make better use of the data deluge. When data science meets education, the numbers previously confined to websites and PDF reports are set free. Teachers, administrators, and consultants apply programming and statistics to prepare data, transform it, visualize it, and analyze it to answer questions that are as unique as are their roles in education.

Data science represents the intersection of domain-level expertise, statistics, and computer programming, applied to data as a means to answer research questions and solve problems. Making sense of data requires, among other things, collaboration, data processing (i.e., wrangling or munging), visualization, and communication of both processes and results in transparent and reproducible ways. Our book focuses on *data science in education*, which we define as using data science techniques like preparing, exploring, visualizing, and modeling data, in order to support schooling at all levels. Our book advances the larger conversation about data science by introducing the idea that the application of data science in a specific field–in this case, education–requires an exploration of unique challenges and the development of unique language. We feel that discussing data science using scenarios that are familiar to education professionals at all levels of education more effectively addresses the needs of those professionals–data analysts and others–who work in that field. This concept of *data science* ***in*** *education* is separate from *data science education*, which seeks to teach the broader techniques of data science while not necessarily teaching the unique application of those techniques in any particular industry. Educators have different needs than data science enthusiasts who aim to self-teach using *data science education* materials. As educational technology transforms both the administrative and student-facing sides of education, it will become increasingly important for education professionals - not just people hired to analyze data - to be able to understand and respond to the data they gather. Our book empowers educators from elementary school to higher education to transform the educational data they are immersed in into actionable insights, in order to help them better serve the students and institutions for whom they work. It could be used as a main textbook in a graduate data science in education course. Alternatively, it could be used as a supplementary textbook for individuals looking to expand their professional toolkit and become more proficient in data science techniques.

.

By the end of this book the reader will understand:

* The diversity of data analysis skills and applications in the education field
* The unique challenges that come with analyzing education data
* That good data analysis has a basic workflow and how they might implement such a workflow
* The wonderful opportunity we have to shape the usefulness of data science in our education jobs

And, the reader will be able to:

* Reflect and determine what their role is as a data analyst within their role as an educator
* Identify and apply solutions to education data’s unique challenges, such as cleaning datasets and working with private student data
* Apply a basic analytic workflow through practice with education datasets
* Be thoughtful, empathetic, and effective when introducing data science techniques in their education jobs

**ABSTRACT**

*Please also provide a concise (150-200 word) and compelling abstract for the book.*

This book aims to give data science in education a common language and set of skills so that the community around this topic can grow and learn more together. It is a friendly welcome to educators who are interested in exploring techniques for data analysis in education. It adds to the growing body of data science books by demonstrating data science using the language, datasets, problems, and workflows that are specific to education. It aims to help educators move from “knowing about” data analysis tools to “knowing how to” conduct meaningful data analyses in their work. This book is ideal for educators at all grade levels looking to start exploring how to use data science in their everyday work, educators who are looking to level up their use of data by learning about programming and machine learning, data-savvy consultants who help educators meet their goals, or experienced data scientists looking for hands on examples that are directly from the education field.

## 2. Detailed Synopsis

The book has three general sections:

1. An overview of what data science in education is and why it is important (i.e., how data science ideas are relevant to education and what challenges are faced by those doing data science in education).
2. An education-focused introduction to a widely-used, freely-available statistical open source software tool, R, and walkthroughs for how R can be used to achieve data and data analysis-related goals.
3. A discussion of what to do once one has a foundation in using data science, including tips on how to be strategic about implementing data science and how educators and education consultants can teach others to use data science tools and methods.

### Outline

Specifically, the outline for the proposed book is as follows:

1. Introduction
   * The Challenge of Data Science in Education
   * Meeting the Challenge
2. How to Use This Book
   * Read the Book Cover to Cover
   * Pick a Chapter That is Useful for Your Level of Experience and Start There\*
   * Read Through the Walkthroughs and Run the Code
3. What is a Data Scientist in Education?
   * Example #1: Leading Office Culture Toward a Data-Driven Approach
   * Example #2: Helping School Districts Plan to Meet Their Goals
   * Example #3: Doing and Empowering Research On Data Scientists in Education
4. The Unique Challenges of Doing Data Science in Education
   * Challenges common to doing data science in any domain
   * Lack of Processes and Procedures
   * Few Guidelines From Research and Evaluation
   * Limited Training and Educational Opportunities
   * The Complex and Messy Nature of Educational Data
   * Ethical and Legal Concerns
   * Analytic Challenges
5. Foundational Skills: Track One (Getting Started)
   * Downloading R and RStudio
     + - To Download R
       - To Download RStudio
     + Check That It Worked
     + Help, I'm Completely New to Using R / RStudio!
     + Creating Projects
     + Packages
       - Installing and Loading Packages
       - Running Functions from Packages
6. Foundational Skills: Track Two (Welcome to the Tidyverse)
   * Loading Data from Various Sources
     + Saving a File from the Web
     + Loading a CSV File
     + Loading Excel files
     + Loading SAV (SPSS) files
     + Loading Google Sheets Files
   * Saving Files
   * Processing Data
     + Selecting Variables
     + Filtering Observations
     + Creating New Variables
   * Communicating / Sharing Results With R Markdown
     + Knit to HTML
     + Knit to PDF
     + Knit to Word
   * Configuring R Studio
   * Getting Data Into and Out of R
     + Clipr
     + Data Pasta
7. Walkthrough #1: Analyzing Gradebook Data Using the Data Science Framework
   * Introduction
     + Background: The ubiquity of gradebook data
     + Purpose of Analysis
     + Summary of Analysis and Data
     + “Big Questions” to be answered
   * Data
     + Secondary Gradebooks
   * Methods
     + Data Wrangling and Descriptive Statistics
   * Analysis
     + Correlation analysis
     + Regression analysis
   * Results
   * Conclusions
   * Appendix
8. Walkthrough #2: Understanding Student Motivation to Learn Using Multi-Level Modeling
   * This walkthrough and subsequent walkthroughs follow the same format was walkthrough #1.
   * This walkthrough explores an increasingly common data type, data output from learning management systems (LMS). Data sources include log-trace measures of how many hours students spent in the course, their final grades, and students’ self-reported interest in and value for the course they were studying.
   * The analysis focuses on an extension of a regression analysis, a multilevel model.
9. Walkthrough #3: Understanding Online Course Motivation with Machine Learning
   * This walkthrough uses the same data set as walkthrough #2, but with a predictive goal, rather than an inferential one.
   * The analysis involves an introduction to machine learning techniques, particularly random forests. It shows methods for assessing the accuracy and effectiveness of the approach and compares the random forest technique to a simple regression approach.
10. Walkthrough #4: Creating Network Graphs of Simulated Data
    * This walkthrough uses relational, or network, data, from participants in an urban STEM teaching graduate certificate program. It includes information on teachers’ self-reported and observed (through Twitter interactions) networks.
    * The analysis includes the steps necessary to process the data (particularly, the steps necessary to create an type of data suited to social network analysis, an edgelist) and to understand selection (e.g., who is choosing to interact with whom?) and influence (how do interactions with influential people impact an outcome, such as teachers’ professional teacher?) processes.
11. Walkthrough #5: Exploring Nationally- and Internationally-Representative Data
    * This walkthrough explores Nationally- and Internationally-representative datasets, with a focus on National Assessment for Educational Progress (NAEP) assessment data.
    * The analysis demonstrates how to access, prepare, and model (in a way that reflects the complex sampling strategy inherent to representative samples) the data to understand differences in students’ achievement across subject areas.
12. Solutions for Adapting Data Science Techniques in Education
    * Matching your analytic strengths to the data needs of your organization
    * The change process
      + Understanding the culture of your organization
      + Building a network
      + To code or not to code
      + A practical guide to introducing new analytic techniques
13. Teaching Data Science
    * A Pedagogical Framework for Data Science Education
      + Problem-based Learning
      + Differentiation
      + Universal Design
    * Adapting Resources From This Book for K-12 Teaching and Learning
    * Tools for Teaching and Learning How to do Data Science
    * Teaching Data Science in the Context of Curricular Standards
14. Resources
    * Books
    * Websites and Web-based Resources
    * Courses and Communities
15. Learning More
    * Work in the Open
    * Welcoming Others
    * Asking For and Receiving Help
    * Adopting a Growth Mindset

### Select References

We provide some select references that we will cite across the above chapters. These references, which we separate by their substantive or methodological focus, will be cited across the text (and so in multiple chapters).

**Substantive References**

Cho, V., & Wayman, J. C. (2014). Districts’ efforts for data use and computer data systems: The role of sensemaking in system use and implementation. Teachers College Record, 116(2), 1-45.

Datnow, A., & Hubbard, L. (2015). Teachers' use of assessment data to inform instruction: Lessons from the past and prospects for the future. *Teachers College Record, 117*(4), 1-26.

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures.* National Academies Press.

Linnenbrink-Garcia, L., Patall, E. A., & Pekrun, R. (2016). *Adaptive motivation and emotion in education: Research and principles for instructional design.* Policy Insights from the Behavioral and Brain Sciences, 3(2), 228-236.

Steinfeld, E., & Maisel, J. (2012). *Universal design: Creating inclusive environments.* John Wiley & Sons.

**Methodological References**

Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using R. Sage publications.

Tatsuoka, M. M., & Lohnes, P. R. (1988). Multivariate analysis: Techniques for educational and psychological research. Macmillan Publishing Co, Inc.

Wickham, H. (2014). *Advanced R*. Chapman and Hall/CRC.

Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. O’Reilly.

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear mixed models: a practical guide using statistical software.* Chapman and Hall/CRC.

### Other Book Features

There is a GitHub repository with all of the code to reproduce the text available here: <https://github.com/data-edu/data-science-in-education>. The repository has already attracted interested--at the time of this writing, it has been ‘starred’ 43 times and is ‘watched’ by 13 GitHub users.

We believe that a unique feature of our book is that we wrote it for educators at all levels–preschool to college. To that end, our writing process includes regularly “testing” content with readers from the education field and receiving feedback about how well we met our content. Their feedback directly informs how we execute on the mission of the book.

A key feature of this book is the pedagogical framework that guides it. As a book on data science in education, we take seriously how the book - itself - will be educative for readers. Our pedagogical framework, then, influences the planning and writing of the book. It consists of three principles: problem-based learning, differentiation, and universal design. We consider each of these in turn.

* **Problem-based learning (PBL)** is a method of instruction that presents learners with a real-world challenge in which they must apply their skills and knowledge to solve. This is critical because we, as authors, do not have all of the right answers in this text: there is not one right statistical model or algorithm, a technique to write code or even software and tools to utilize. Thus, the text features continuing challenges in the form of the five walkthroughs that reflect the types of challenges that educational data scientists may encounter in the course of their work. At the end of each tutorial, there exist exercises that provide the opportunity for readers to extend and apply the ideas presented in the chapter. Moreover, the challenges are structured in such a way that readers return to some of them, but with different aims, over the course of the book.
* **Differentiation** is a method for providing multiple pathways for learners to engage with, understand, and ultimately apply new content and new skills. We consider inclusivity (in terms of who belongs as a part of the audience for this text and how this broader view of who participates in data science implies the types of challenges, topics, and data that we include in the book) and accessibility (technically, in terms of how a wide audience of readers is able to access and use the book, as well as in terms of the ways in which the content is written based on the unique assets that those in education bring) along with how we differentiate the book. We differentiate the book by recognizing and providing background content/skills (either explicitly or through reference to other resources), embedded checks for understanding, and recommendations for where to begin based upon prior expertise. We also provide links to videos of coding and screenshots that are annotated and reflective of the content in the text to help to show readers how they are able to use what they are reading about.
* **Universal Design** is a series of principles which guide the creation of spaces that are inclusive and accessible to individuals from all walks of life regardless of age, size, ability, or disability. While traditionally applied to physical spaces, we have extended these principles to the creation of a data science text in such a way that the text and accompanying materials will be designed for individuals from all walks of life, regardless of educational level, background, ability, or disability. Many of the seven guiding principles of Universal Design are readily transferable to the creation of a text, such as equitable use, flexibility in use (aided in large part through differentiation), simple and intuitive use, perceptible information, and tolerance for error.

## 3. Description of the Target Market

The primary audience for this book are educators and data practitioners in education who have varying levels of experience with data science. They are administrators, researchers, data analysts, teachers, or college students studying in an education-related field such as teaching or academia. These individuals may also work for non-profit and governmental organizations that take a data-informed approach to helping schools meet their goals. Though they may play a variety of education roles and work in a variety of education levels, our audience has something in common: they seek to make the best learning experience possible for students.

Though as a whole our audience has varying levels of data science experience, they all have one thing in common: they are excited to use data in their current or future education-related career. They may use spreadsheet tools like Excel or Google Sheets, or they may be exploring programming in R as a way to advance their use of data. In addition, it is well established that school districts and charter networks are utilizing data to drive decision-making at an increasing rate. Given the greater use of data in education, the audience may also be current users of R who are looking for examples of data science work that are applicable to an education setting.

We can point to some data points that suggest interest in this topic. First, most graduate students in educational research take one or more research methods courses. While there have been a few examples of such courses involving the use of R, they are few and far between, and there is not an introductory text (yet) available for such classes. The nascent interest we have seen in our book (as evidenced by the interest on GitHub; see the Other Book Features section for details) suggests an interest on the part of students and early career research faculty in education who wish to learn R but do not have resources readily available to help advance their efforts.

Second, there are 25,000 members in the largest American educational research organization--the American Educational Research Association (AERA). Members of this organization and the individuals the organization represents (graduate students, faculty members at universities, and researchers working in organizations such as Educational Testing Service) make up parts of our book’s audience. In fact, 7 of 28 professional development workshops at the most recent AERA conference included the use of R.

Lastly, there is a diverse audience that makes use of education data within each of the 13,500 school districts in the United States. This audience includes parents, teachers, service providers, and researchers[[1]](#footnote-0). With this many school districts in the US alone, there is an audience of data practitioners in education who would benefit from developing fluency in using R as an integral component of their analytical workflow.

## 4. Competing Titles

A search for “data science in education” on Amazon does not return results that have the same mission as our work – to create a common language for using data science in the education field and to teach data science using datasets that are familiar to educators. However, there are some books that teach data science in academia and some books that teach data-informed decision making in the education field.

1. *Data Science in Higher Education: A Step-by-Step Introduction to Machine Learning for Institutional Researchers* (Lawson). Our book aims to target an audience that is wider than only higher education researchers. We aim to reach teachers and administrators who work in schools as well as those who train them.

2. *The Educational Leader's Guide to Improvement Science: Data, Design and Cases for Reflection (*Crow et al.). Improvement science is an example of an increasingly popular way that data analysis is being used in the education field. However, most books on improvement science do not teach tools for rapid and large scale analysis during the “plan, do, study, act” cycles commonly used in the improvement science model. Our book aims to provide depth in this area so that educators are armed with the technical tools needed to evaluate if their change ideas are truly helping students.

3. *Quantitative Ethnography* (Schaffer). Instead of addressing the needs of university-level researchers, as this book is, our book is aimed at users of data science who need examples of applying programming and statistics on data commonly found in the broader education field. Further, our book aims to focus less on using a level of statistical methods required for scientific journals and instead focuses on a standard of methods that allows for rapid but sound analytic results for educators.

4. *R for Data Science* (Wickham et al.). This book features the use of R from a contemporary perspective, one that emphasizes the use of the tidyverse suite of packages for which Wickham and colleagues (at RStudio and elsewhere) have become widely known. This book has quickly come to be recognized as one of the most useful texts for learning to use R for data science. However, this book does not focus on the data types, needs, and potential assets that those working in education bring to data science.

*5. Data Science From Scratch* (Grus). This book takes a pedagogical approach to data science by describing data science methods in terms of their underlying, constitutive parts. For example, this book explains correlations in terms of how, computationally, two functions to calculate the variance and covariance can then be composed together to find how two variables are related. This book is related to ours because it aims to teach data science. It is different, however, in that it is focused on the use of Python - rather than R - and because it does not speak, particularly, to educators and educational data practitioners.

## 5. Format and Timeline

* The book’s approximate length will be 200 pages
* Illustrations or screenshots will be used
* There are no special considerations that apply to our plans for the book, including unusual format, use of color, hard-to-get illustrations, or anything else calling for unusual resources
* The anticipated delivery schedule is as follows:
  + Two draft chapters to be delivered by: June 30, 2019
  + Half draft manuscript to be delivered by: August 31, 2019
  + Full draft manuscript ready for tech review delivered by: October 31, 2019
  + Final and full manuscript ready for production: December 31, 2019

1. <https://2pido73em67o3eytaq1cp8au-wpengine.netdna-ssl.com/wp-content/uploads/2016/03/Who-Uses-Student-Data-Infographic.pdf> [↑](#footnote-ref-0)