# Learning Analytics Goes to School

## A Collaborative Approach to Improving Education

Andrew Krumm
Barbara Means
and Marie Bienkowski

ROUTLEDGE

# Learning Analytics Goes to School

*Learning Analytics Goes to School* presents a framework for engaging in education research and improving education practice through the use of newly available data sources and analytical approaches. The application of data-intensive research techniques to understanding and improving learning environments has been growing at a rapid pace. In this book, three leading researchers convey lessons from their own experiences—and the current state of the art in educational data mining and learning analytics more generally—by providing an explicit set of tools and processes for engaging in collaborative data-intensive improvement.

**Dr. Andrew Krumm** is Director of Learning Analytics Research at Digital Promise, a nonprofit organization that brings together the expertise of educators, researchers, and technology developers in the interest of improving teaching and learning. Dr. Krumm has launched multiple research-practice partnerships and his research addresses the use of data-intensive research techniques to improve learning environments.

**Dr. Barbara Means** is Executive Director for Learning Sciences Research at Digital Promise. Formerly the founder and director of the Center for Technology in Learning at SRI International, Dr. Means is a nationally recognized expert in defining issues and approaches for evaluating the implementation and efficacy of technology-supported educational innovations.

**Dr. Marie Bienkowski** is Director of the Center for Technology in Learning at SRI International, a nonprofit research and development organization based in Silicon Valley that takes innovative ideas and technologies from the laboratory to the end-user and marketplace. Dr. Bienkowski is a computer scientist and education researcher leading efforts to improve student learning, effective teaching, and meaningful assessment.

# Learning Analytics Goes to School

## A Collaborative Approach to Improving Education

Andrew Krumm, Barbara Means, and Marie Bienkowski

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2018
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

*Routledge is an imprint of the Taylor & Francis Group, an informa business*

*Library of Congress Cataloging-in-Publication Data*
Names: Krumm, Andrew, author. | Means, Barbara, 1949– author. | Bienkowski, Marie, author.
Title: Learning analytics goes to school : a collaborative approach to improving education / Andrew Krumm, Barbara Means, and Marie Bienkowski.
Description: New York, NY : Routledge, 2018. | Includes bibliographical references.
Identifiers: LCCN 2017040312 (print) | LCCN 2017059502 (ebook) | ISBN 9781315650722 (e-book) | ISBN 9781138121829 (hardback) | ISBN 9781138121836 (pbk.) | ISBN 9781315650722 (ebk)
Subjects: LCSH: Education—Research—Data processing. | Educational statistics. | Big data.
Classification: LCC LB1028.43 (ebook) | LCC LB1028.43 .K78 2018 (print) | DDC 370.72—dc23
LC record available at https://lccn.loc.gov/2017040312

ISBN: 978-1-138-12182-9 (hbk)
ISBN: 978-1-138-12183-6 (pbk)
ISBN: 978-1-315-65072-2 (ebk)

Typeset in Sabon
by Apex CoVantage, LLC

# Contents

# Figures

# Tables

# Boxes

# Preface

This book is the product of a collaboration among three researchers—Andy, Barbara, and Marie—who were supported by the insights and efforts of many colleagues and partners. The building blocks of this book stretch back multiple years, but the catalyst for writing it came from the rapidly developing set of lessons we were learning from research partnerships that involved using data-intensive research techniques in collaboration with educational practitioners.

Starting in 2012, the three us began wrestling with what it means to use new forms of evidence in the service of educational improvement. Barbara, for example, was working to understand the landscape of how to use data stemming from digital learning environments and administrative data systems through her brief sponsored by the U.S. Department of Education on *Expanding Evidence Approaches for Learning in a Digital World*. Andy and colleagues began a project on measuring learning behaviors and strategies using data from digital learning environments. This project involved working with both researchers and practitioners to better understand how measures of learning behaviors could be shared and used to improve digital and face-to-face learning environments. Over time, this project led to a partnership with the Carnegie Foundation for Advancement of Teaching and the Carnegie Math Pathways, a networked improvement community working to solve the developmental math crisis in the United States.

The partnership with the Carnegie Math Pathways presented us with a unique opportunity to learn first-hand the principles, practices, and tools of improvement science. Subsequently, we began integrating improvement science approaches into more and more of our data-intensive research work. Throughout this book, our experiences working with the Carnegie Math Pathways serve as one of two anchoring cases illustrating necessary conditions for engaging in partnership-driven, data-intensive improvement research.

Prior to our partnership with Carnegie, Andy participated in a unique event hosted by the National Science Foundation referred to as an Ideas

Lab. The premise of the Ideas Lab was to bring researchers together for a concentrated amount of time to promote the development of collaborations that could solve pressing problems related to data-intensive research in education. Andy worked with Anna Gassman-Pines from Duke University at the event and formed a collaboration around merging data from two statewide agencies in North Carolina. Andy and Alex Bowers from Teachers College, Columbia University also met at the Ideas Lab, and along with Mingyu Feng and a successful charter management organization—Summit Public Schools—formed a research-practice partnership to identify necessary conditions for engaging in collaborative data-intensive research. Our partnership with Summit represents the second anchoring case for the book.

As these and other partnerships were developing, Andy, Barbara, and Marie collaborated to identify ways of harnessing the technical capabilities of various research labs within SRI International—our home organization for much of our time working together—to apply new machine learning techniques to data originating from a variety of digital learning environments. We sought out experts in other parts of SRI who could assess the viability of using machine learning to develop data products that could directly help practitioners. Through these experiences, we quickly came to appreciate the importance of wrapping data-intensive research techniques within an overarching process that includes deeply understanding the problems to be solved from the perspective of educational practitioners and drawing on what can be learned from prior research before diving into a dataset from a digital learning environment or administrative data system. Similarly, we realized the importance of working with practitioners to jointly interpret a data analysis as well as co-develop potential future courses of action inspired by an analysis.

Throughout these varied experiences, we have learned a lot about what does and does not work in using data-intensive research methods to improve learning environments. Our goal in this book is to convey lessons from our own experiences as well as the current state of the art in the field of educational data mining and learning analytics in the context of an explicit set of tools and processes for engaging in collaborative data-intensive improvement.

As researchers with diverse backgrounds, we share a commitment to using research evidence to inform educational policy and practice and an enthusiasm for engaging in partnerships with educators to pursue data-intensive research. We came to this place by different paths, however, reflecting our individual methods of training and research experiences. Hopefully understanding our different paths will help newcomers to the field see ways of entering into the exciting work taking place at the intersection of data-intensive research, educational improvement, and the learning sciences.

After earning her Ph.D. in computer science at the University of Connecticut, Marie came to SRI to join the Applied Artificial Intelligence Program, where she worked on intelligent tutoring systems and applied methods for analyzing learners' explanations of their thinking to military training. The desire to apply her skills to K–12 education led Marie to transfer to the Center for Technology in Learning, where she started working with Barbara.

Barbara earned her doctorate in educational psychology at the University of California, Berkeley, where she conducted experimental studies on children's learning and memory. A desire to move out of the laboratory and into research in real-world education settings eventually brought her to SRI, where she founded the Center for Technology in Learning in 1993. Her research experiences include numerous studies of how learning technologies are implemented in schools and one of the first studies of how teachers and school leaders use data from web-based data systems to inform school improvement efforts.

Andy earned his doctorate in learning technologies at the University of Michigan, where he studied the diffusion and implementation of technologies, data use in schools, and was an early contributor to the University's developing learning analytics practice. As a researcher at SRI and Digital Promise, Andy has worked with data from dozens of learning technologies and has supported multiple research-practice partnerships.

# Acknowledgements

This book is the direct result of multiple years of work from multiple collaborators. At an organizational level, we wish to thank SRI International, the Carnegie Foundation for the Advancement of Teaching, and Summit Public Schools. Within SRI International, many individuals have had a hand in shaping the work described in this book, including Mohamed Amer, Jared Boyce, Cynthia D'Angelo, Ajay Divakaran, Mingyu Feng, Larry Gallagher (Stanford University), Jim Klo, Bladamir Lopez-Prado, Timothy Podkul, Behjat Siddiquie (Amazon Go), Erica Snow (Imbellus), and Tallie Wetzel. Alex Bowers and colleagues at Teachers College, Columbia University were instrumental in applying hierarchical cluster analysis tools and in developing school-based models for engaging in educational data science. Anna Gassman-Pines and colleagues at Duke University were instrumental in opening up the full complexities of multi-institutional statewide partnerships. Adam Carter, Jon Deane (Chan-Zuckerberg Initiative), Ross Lipstein, Vishal Shah (College Track), and Howard Shen and were key collaborators at Summit Public Schools. Rachel Beattie, Ann Edwards, Heather Hough (Policy Analysis for California Education), Amanda Meyer (Policy Analysis for California Education), Jojo Manai, Bryan Sulcer, Chris Thorn, and Hiroyuki Yamada were key collaborators from the Carnegie Foundation for the Advancement of Teaching and Carnegie Math Pathways.

This work was supported by the National Science Foundation (SMA-1338487, DRL-1444621, and DRL-1418332), the William and Flora Hewlett Foundation, and SRI International. The opinions expressed in this book are those of the authors and do not necessarily represent those of the above organizations.

Chapter 1

# Introduction

The daily activities of schools and universities—from taking attendance to assessing students—can leave a trail of data that, under the right conditions, can be used to explore teaching and learning like never before. Until recently, though, researchers had to choose between collecting rich data on a small number of individuals or amassing less rich data for larger numbers of individuals. And in both cases, collecting data on the same individuals over time required significant costs and complexities. For activities that take place in digital learning environments like games, learning management systems, and intelligent tutoring systems, surprisingly rich data can be collected on dizzyingly large numbers of learners over time. While opportunities to collect and analyze new forms of data increase every day, critical challenges need to be overcome in order to use these data to improve teaching and learning.

Along with new forms of data, such as system log data (i.e., records of users' interactions with a digital learning environment), familiar forms like text, audio, and video are becoming increasingly open to in-depth analysis—at scale—through machine learning and artificial intelligence. These newly found and newly analyzable data are often described as "big data" both inside and outside of education. Over the past decade, analyzing educational big data has largely occurred in research labs at universities, technology companies, and non-profit research institutes, and this basic research, with few exceptions, has yet to diffuse widely or to fundamentally change teaching and learning (Baker, 2016; Martin & Sherin, 2013). Where there have been successes, such as with the ASSISTments platform (Roschelle, Feng, Murphy, & Mason, 2016) and in examples described later on in this chapter, new forms of data and new analytical techniques have been grounded in problems facing practitioners and used to develop and assess potential changes related to those problems.

As some have argued, improving teaching and learning at scale will require new ways of organizing the work of educational research (Bryk, Gomez, Grunow, & LeMahieu, 2015). Starting around the same time as educational data mining and learning analytics—some of the most recognizable fields in what may be termed data-intensive research—an approach to conducting educational research referred to as research-practice partnerships was taking shape (Coburn, Penuel, & Geil, 2013). While the idea of forming partnerships is not new, frustrations with the status quo, a critical mass of success stories, and new funding opportunities have coalesced into an overarching approach where researchers work on pressing problems of practice in an iterative and collaborative fashion with practitioners (Penuel & Gallagher, 2017). In many ways, researchers working under the banner of research-practice partnerships have found a way to directly impact teaching and learning—by working directly with teachers and learners. While a disarmingly simple idea, this approach has profound implications both for *who* participates in the work of improving learning environments and for *how* that work is carried out.

In this book, we describe multiple efforts to use data-intensive research methods to improve teaching and learning. In particular, we highlight the important role that partnerships between researchers and practitioners can play in activating educational big data as a resource for improvement. Through the lens of what we refer to as *collaborative data-intensive improvement* (CDI), we aim to make explicit the ways in which educational researchers can engage in longer-term partnerships with the goal of not just understanding learning but also of improving outcomes in real-world learning environments. Doing this well, we believe, will require a fundamental rethinking of how data are used for research and improving practice.

## Data-Intensive Research in Education

This book offers an introduction to the developing fields of educational data mining and learning analytics by describing goals, methods, and examples. In outlining the past, present, and potential future for these fields, throughout this book, we focus our descriptions on using data and complex data analyses to improve learning experiences and educational outcomes. We illustrate this potential with firsthand examples that span multiple academic content areas, learning environments, and learner types. We provide examples of decision making at the classroom, school, and education system levels taken from schools, universities, and community colleges.

Along with examples from our own work, we will describe how other researchers have employed educational data mining and learning analytics to address problems that originate in one form or another from the front lines of teaching and learning. In describing multiple examples and analytical approaches, we will highlight potential benefits and costs associated with each. The reader should know, however, that we are not attempting to provide a balanced treatment of all approaches. Our emphasis will be on *collaborative* data-intensive research approaches that prioritize shaping practical improvements over advancing analytic methods. While we will not restrict our coverage solely to collaborative data analysis approaches, they will be our lens for choosing what to highlight in a rapidly changing landscape. We hope that both researchers and practitioners will find this lens useful in making sense of new sources of education data, new analytic techniques, and new opportunities to form partnerships.

## *The Challenge of Jargon*

One challenge facing newcomers to the field of data-intensive research is the wave of jargon they are likely to encounter. Already, in the first few pages of this book, we have referred to educational data mining, learning analytics, system log data, and big data. In an attempt to keep jargon to a manageable level, we have made explicit choices about the terminology we use in this volume, recognizing that some key details, distinctions, and research histories will be lost in this translation.

Before progressing further, we would like to orient the reader to a few key terms: educational data mining, learning analytics, data-intensive research, and educational data scientist.

*Educational data mining* and *learning analytics* represent distinct fields that have a high degree of overlap (Siemens & Baker, 2012). For simplicity, and to contrast these fields with other research traditions, we will refer to both of them as examples of *data-intensive research in education*. The additional fields that we want to integrate into learning analytics and educational data mining include studies of data use in schools (e.g., data-driven decision making) and collaborative research approaches (e.g., design-based implementation research and improvement science). As we will describe in Chapter 5, these additional fields are important both to the past and to the present of data-intensive research in the same ways that learning analytics and educational data mining are.

Educational data mining, which predates the field of learning analytics, largely concentrates on using machine learning techniques to identify patterns within large educational datasets, often from specific digital learning environments like intelligent tutoring systems. Oftentimes, these same technologies are what deliver interventions aimed at improving learning. Learning analytics, on the other hand, tends to focus less on machine learning techniques and more on statistical and visualization approaches, whereby interventions aimed at improving learning are delivered as much by an individual as a technology. As Baker and Inventado (2014) point out, the differences between these two fields grew out of different interests and backgrounds of the researchers in the two areas, and do not reflect any fundamentally opposing beliefs about how people learn. They agree on the assumption that data collected at scale and analyzed with rigorous methods will help arbitrate between different theories and proposed practices (Bienkowski, Feng, & Means, 2012).

Data-intensive research "involves data resources that are beyond the storage requirements, computational intensiveness, or complexity that is currently typical of the research field" (Dede, 2015, p. 2). The field of education more generally is gradually expanding its data repertoire to include data from digital learning environments and from increasingly sophisticated administrative data systems. In addition, other familiar forms of data, such as video and audio files, can now be explored at scale with greater speed. Therefore, we use the term data-intensive research to integrate these developing examples as well as those stemming from educational data mining and learning analytics.

An *educational data scientist* is someone who practices data-intensive research in education. The term "data scientist" is expansive and touches on multiple knowledge, skills, and abilities (see O'Neil & Schutt, 2013). Anyone who uses data-intensive research methods is often referred to as a data scientist. And while data science has become a hot new career (Ferenstein, 2016), the knowledge, skills, and abilities needed to perform this role are often ill-defined, especially in education (Piety, Hickey, & Bishop, 2014). Generally speaking, a data scientist is an individual with some combination of computer science skills, a background in statistics and mathematics, and relevant domain expertise (O'Neil & Schutt, 2013). Agasisti and Bowers (in press) define an educational data scientist as an individual who has "the technical skills to collect, analyze, and use quantitative data, and at the same time the managerial and communication skills to interact with decision-makers and managers at the school level to individuate good ways of using information in the practical way of improving practices and initiatives" (p. 6). In the coming chapters, we elaborate on these descriptions and make the case that an educational data scientist is someone who clarifies how data-intensive research methods can be used to address questions of importance to educators, carry out the actual analyses, and help develop and refine ideas for improvement.

## Focus of the Book

Given the continuing proliferation of data and the increased sophistication of data-intensive research techniques, now is a good time to take stock of data-intensive research in education, articulate fruitful directions for advancing the field, and provide an onramp for newcomers. In working to achieve this ambitious and multifaceted aim, it is important to clarify what this book will and will not deliver. First, this book is not a how-to guide on data-intensive research methods in education. The interested reader can explore a growing number of learning analytics focused Massive Online Open Courses (MOOCs) for this purpose, such as Ryan Baker's *Big Data and Education*, Tim McKay's *Practical Learning Analytics*, and the University of Texas at Arlington's upcoming MicroMasters on Learning Analytics. In addition to educational applications of analytics, a researcher or data scientist, at some point, will need to group rows of data and apply a function, such as identifying the average amount of time a student spent in a digital learning environment across multiple sessions. Depending upon one's chosen software package, without too much difficulty, one could use a search engine to identify a serviceable answer. Less searchable are strategies for identifying sources of data in the first place and knowing how to work with practitioners to apply the right analytical technique to the right data and how to structure a meeting where researchers and practitioners come together to interpret and draw implications from a data-intensive analysis. In many ways, that is what this book is about.

This book is also not a standard course in educational research design or a program in educational leadership, though it does include elements and insights from these fields. It presents some fundamental research and leadership concepts as they relate to each other and to the goal of using data-intensive research to improve education outcomes. We seek to equip readers with an understanding of methods to enable clearer thinking about how new sources of data and new analytical techniques could help them create more desirable outcomes for students.

## Examples of Data-Intensive Improvement

When Romero and Ventura (2007) surveyed the data mining literature for education applications published between 1995 and 2005, they found only two articles published before 2000. In contrast, by 2016, a Google Scholar search returned over *one million articles* on this topic. And educational applications of data-intensive research have moved beyond scholarly publications to capture the public's imagination through popular press coverage such as a recent *New York Times* article, "Will You Graduate? Ask Big Data" (Treaster, 2017). In the following sections, we describe three diverse examples to introduce some of the possibilities.

### *Measuring Chronic Absenteeism and Its Causes*

School districts have always kept data on their students, but it used to be hard to access or to organize the data in a way that would shed light on educational issues. For example, a school district would have a record system showing the number of students in attendance each day and would routinely compute the average daily attendance for the school year. Schools with attendance average daily rates over 90 percent generally believed they were doing very well on this metric. But most schools, districts, and states did not have the capability to look at attendance patterns for *individual students* over multiple years or to relate students' attendance patterns to their educational outcomes (Balfanz & Byrnes, 2012). Without such a longitudinal student-level dataset, schools were missing the story of what has come to be called "chronic absenteeism"—missing 10 percent or more of school days in an academic year.

The importance of attending school has long been recognized, but until recently we lacked the ability to quantify the impact of chronic absenteeism on educational outcomes and hence any basis for saying what level of absenteeism should be cause for concern. Increased computing capacity, improved tools for bringing together data from different data systems, and the use of unique, statewide student identification numbers permitting linking multiple student-level datasets have enabled exploration of the issue of absenteeism in states and districts.

Analysis of data from Chicago Public Schools (CPS) by researchers from the University of Chicago Consortium on School Research, for example, found that missing 10 or more days of school during the year, whether excused or not, was a stronger predictor of school failure than low test scores at the end of the prior school year. Analysis of the CPS data also showed that ninth graders with high test scores who missed two or more weeks of school were more likely to fail than students with low test scores who were absent five or fewer days (Allensworth & Easton, 2007). An issue brief from the University of Chicago's To&Through project indicated that each week a student is absent during a semester of ninth grade is associated with a 20 percent decline in the probability of earning a high school diploma. After becoming aware of the data on chronic absenteeism and its correlates, CPS began implementing a number of programs to address chronic absenteeism. One strategy involved improving the accuracy and availability of individual students' attendance records so that teachers and school leaders would be motivated to examine them on a weekly basis in order to identify students in need of intervention. Another strategy involved creating a culture of collective responsibility around attendance. Some schools started talking about the importance of attendance at school assemblies and posted attendance charts in school hallways. The combination of these and other approaches over the past decade have led to a 17 percent increase in high school graduation rates (To&Through Project, no date).

### *Using Learning Analytics to Improve Digital Learning Systems*

System log data from digital learning environments are particularly promising because they can capture *who* did *what* and *when*. Researchers and educational data scientists can explore this kind of data to look at the sequences of actions taken by individual learners, greatly expanding the potential to examine detailed learning activity data at a massive scale in order to glean insights into the *processes* of learning. Being able to go beyond analysis of outcomes to delve into learning processes opens up significant opportunities for improving both digital and face-to-face learning environments.

Since the 1980s, Carnegie Mellon University (CMU) has pioneered the design, development, and evaluation of digital learning systems that employ learning theory and artificial intelligence to adapt to the responses of individual learners (Koedinger & Corbett, 2006). More recently, with the availability of increased data storage and analysis capabilities, researchers at CMU began applying a variety of machine learning and statistical techniques to the data produced when students use their tutoring systems in order to derive insights into how to improve those tutoring systems (Koedinger, Stamper, McLaughlin, & Nixon, 2013).

A hallmark of the tutoring systems developed at CMU is that they are based on a detailed cognitive analysis of the knowledge and skill components needed in the domain being studied. Each problem presented in the tutoring system was designed to assess one or more knowledge components. One of the types of data researchers extract from the tutoring system's log files is whether the learner made an error or answered correctly each problem involving a given knowledge component. Ken Koedinger and Elizabeth McLaughlin of CMU leveraged this kind of data in a recent study in which middle school students solved large numbers of beginning algebra problems online, including the three problem types shown in Table 1.1. The target proficiency in this study was being able to solve two-step story problems, such as the one shown in the left-hand column. The researchers wanted to figure out what kind of practice would best support students in acquiring this skill.

*Table 1.1* Story Problem Types Studied by Koedinger and McLaughlin

| | *Problem Type* | |
|---|---|---|
| *2-step Story Problem* | *1-step Story Problem* | *Substitution Problem* |
| Ms. Lindquist is a math teacher. Ms. Lindquist teaches 62 girls. Ms. Lindquist teaches *f fewer boys* than girls. Write an expression for how many students Ms. Lindquist teaches. | Ms. Lindquist is a math teacher. Ms. Lindquist teaches 62 girls. Ms. Lindquist teaches *b boys*. Write an expression for how many students Ms. Lindquist teaches. | Substitute 62-f for b in 62+b<br>Write the resulting expression. |
| Answer: 62+62-f | Answer: 62+b | Answer: 62+62-f |

Source: Koedinger and McLaughlin (2016).

Many instructional designers and educators would hypothesize that practicing one-step problems to mastery would be the best foundation for moving to the harder two-step problems. Although this makes intuitive sense, a major theoretical assumption in the CMU work is that difficulty levels predict transfer because both are a function of the same underlying required knowledge components. When the Carnegie Mellon team analyzed data from the log files for these three types of problems, they found that students had more difficulty with substitution problems, like that in the third column, than they did with one-step story problems. For this reason, Koedinger and McLaughlin predicted that it would be more beneficial to practice symbolizing algebraic terms in the substitution problems than to practice one-step story problems.

Using a web-based tutoring system, the researchers randomly assigned 711 middle school math students to either substitution practice or one-step story problem practice in preparation for two-step story problems. Findings supported the researchers' hypothesis that production of symbolic representations was the key prerequisite for learning to solve the two-step algebra problems. Prior practice on substitution problems based on the cognitive model generated from this data-driven approach inspired an intervention that subsequent experimental testing showed would enhance learning on the target skill of two-step algebra word problems.

The researchers interpret this finding as support for their assumption that task difficulty data can be used as a proxy for skill transfer data. They point out the practical significance of this finding: Direct testing of the transfer of skills from one type of problem to another requires setting up an experiment to test performance on task B with and without prior practice on task A. Generating task difficulty parameters automatically through data mining can provide the input needed for cognitive models so that instructional design and development work can proceed more quickly and more ethically (Koedinger & McLaughlin, 2016).

### *Identifying College Students at Risk of Dropping Out*

Our third example returns to the challenge of identifying students at risk of leaving school, but in this case at the college level. Earning a college degree has major consequences for employability and lifetime earning (Pascarella & Terenzini, 2005). Thanks to the wide range of higher education options in the U.S., including institutions with open admissions, increasing proportions of young people from all backgrounds start some kind of college program. But *completing* a college program with a degree or industry-recognized credential is something different. For students entering college for the first time in 2009, for example, only 53 percent earned a bachelor's degree by 2015, six years later.

As state and federal governments have increased their scrutiny of completion rates for individual colleges, those institutions have become acutely aware of the need to increase the proportion of their students who are retained from year to year and actually leave with a degree. Colleges and universities have turned to data-intensive research techniques to help them identify students who are at risk of failing to complete a course or program of study. Measuring graduation rates requires connecting the academic records from different terms for each individual student to measure whether that student persisted from one term to the next. By combining data from admissions applications and transcripts with data on performance in a particular course, analysts found they could identify groups of students at risk so that those students' instructors or academic advisors could work with them to avoid course failure and dropping out (Hanover Research, 2014).

Tim Renick, Vice President for Enrollment and Student Success at Georgia State University, describes a well-known case of using data-intensive approaches to enhance college completion rates (Renick, 2017). In 2003 this urban public university saw just 33 percent of white students who had enrolled as freshmen and just 22 percent of under-represented minorities who had enrolled as freshmen leave the college with a bachelor's degree. By 2017, Georgia State's degree completion rate had risen to 65 percent for both groups of students, making Georgia State the only public university in the nation where the completion rate for under-represented students is equivalent to that for white students.

Georgia State implemented multiple changes in its practices and interventions with students at risk to achieve these results (Kurzweil & Wu, 2015), but a key enabler was a collaboration with EAB (formerly the Education Advisory Board). EAB helped Georgia State comb through 10 years of student data records—over 2.5 million course grades. These analyses provided insights such as the fact that prospective political science majors who got an A or B in their first political science course had a 75 percent probability of graduating on time, while those who got a C had only a 25 percent probability of doing so (evoLLution, 2016). The university had been doing nothing to follow up with students who earned Cs in their gateway courses because a C grade is adequate to earn the course credit toward graduation. Georgia State hired more academic advisors in order to act more promptly on information identified by an analysis.

The Graduation and Progression System (GPS) academic advising dashboard developed by Georgia State and EAB displays real-time analyses of students' academic progress and the implications of certain decisions, such as taking courses out of the usual sequence.

The GPS displays results from a system that tracks students for 800 different alerts that can trigger action:

> Now, every day the system searches all of our student-information systems for evidence of any of these 800 things. Did a student register

> for the wrong course? Did they do poorly in a prerequisite course? Are they in a major that does not fit their ability? When an alert goes off, an advisor proactively reaches out to the student, typically within 48 hours.
>
> (evoLLution, 2016)

Higher education institutions are also starting to combine the relatively stable information from academic records with more timely information from their campus learning management systems. Learning management systems (LMSs) are online systems that support instructors in delivering course content and assessments; many LMSs include interactive features such as discussion boards. Measures such as the number of days on which a student logs in to the LMS compared with other students in their class, scores earned on interim assessments within the LMS, engagement with course materials, and participation on discussion boards, all measured relative to other students in the same course, can in certain circumstances be used to predict likelihood of completing the course. Combining LMS data with other types of data, the firm Civitas Learning has helped several of its client institutions identify individuals among their high-GPA students who were showing signs of disengagement with college. These high-achieving disengaged students had tended to fall through the cracks because they did not have obvious markers of course failure in their academic records.

Common across all the examples cited previously is the use of data-intensive research methods for identifying and working to improve educational processes. In all cases, data helped identify an opportunity to improve but the data didn't solve the problem—that was up to teams of people in each education institution.

### *Why Engage in Collaborative Data-Intensive Improvement?*

In the examples cited previously, the work of identifying problems to solve, collecting and analyzing data, and deriving implications is prototypical of the work of researchers and educational data scientists engaging in a style of inquiry conducted in concert with educators that we refer to as *collaborative data-intensive improvement* (CDI).

As the use of digital learning environments in schools increases and more and more data are captured in administrative data systems, researchers and data scientists who can support CDI may increasingly be called upon to not only extract meaning from data but also to structure specific activities before and after developing data products. These before and after activities are critical, as they help partnerships translate what is learned from a data-intensive analysis into specific actions that can be used to solve local problems of practice (Krumm, Waddington, Teasley, & Lonn, 2014).

Along with the growing evidence for the benefit of combining data-intensive research with specific efforts to improve teaching and learning, three trends give us optimism that the time is ripe for engaging in data-intensive research in education, and in particular, CDI:

- ***Lessons learned from the data driven decision-making movement.*** For decades, schools have been pressed to use data to drive their instructional and organizational decision making. Multiple scholars have examined what worked and what didn't from this period, and as a field, we are moving beyond viewing data as inherently actionable or as a self-activating resource.
- ***Increased role and importance of research–practice partnerships.*** Both private philanthropies and federal agencies are supporting this trend by providing funding for collaborations between researchers and practitioners. Pioneering efforts from organizations, such as the Carnegie Foundation for the Advancement of Teaching, are providing useful models for how these partnerships can work (Coburn & Penuel, 2016).
- ***The availability of data and the need to interpret them responsibly.*** Data of increasing size and variety are available as never before. With this growing resource will come a need to structure appropriate analyses, draw appropriate conclusions, and structure follow-on activities.

Engaging in CDI opens up unique possibilities for researchers and educational data scientists; education leaders and practitioners; and technology developers.

**For researchers and educational data scientists** who want to see their work improve the quality of education and the equity of opportunities that students receive, collaborative forms of data-intensive research offer opportunities to directly experience and participate in the improvement process. Researchers are accustomed to publishing their analyses and research conclusions in technical reports and scholarly journal articles, which are often not read by the education decision makers and practitioners responsible for the educational experiences that students actually receive. Even when a study does get wide publicity in the general press or in venues where educators gather information, such as their professional conferences or periodicals, a research report is not self-explanatory; understanding how to apply an insight from research to a new context is challenging for researchers and practitioners alike. By directly engaging with practitioners in a partnership, researchers have the opportunity to see their work put to use in real learning environments in ways they believe are well reasoned and likely to be successful.

**For education leaders**, engaging in data-intensive work with researchers offers an opportunity to increase the likelihood of ameliorating an important problem of practice by applying a systematic set of tools and approaches and enlisting additional intellectual resources in the form of researchers and data scientists. The magnitude of improvement that is possible has been demonstrated by the Carnegie Math Pathways from the Carnegie Foundation for the Advancement of Teaching, which we will describe throughout this book. In some cases, university systems and school systems eager to apply data-intensive approaches to their improvement efforts are funding the work of their external collaborators directly (Treaster, 2017), but in other cases researchers have their own funding to support their participation (UT Arlington News Center, 2014).

**For teachers and instructors**, CDI complements a commitment to the scholarship of teaching and learning. This form of scholarship entails reflective inquiry into student learning in specific academic domains and seeks to generate insights that improve teaching and thereby enhance student learning (Hutchings, Taylor Huber, & Ciccone, 2011). Lee Shulman, one of the early advocates for this form of inquiry, offers three rationales for this kind of work that are equally applicable to CDI (Shulman, 2000). First, there is professionalism, which Shulman describes as the "inherent obligation" entailed in being a professional educator and in representing the discipline one teaches. Second, there is the pragmatic rationale: An educator should strive to make sure that his or her work is constantly improving and enabling students to meet their learning goals. Finally, there is the need to be able to demonstrate to external authorities such as administrators, school boards, and accrediting agencies that one's teaching is adding value for students and improving over time. CDI can enhance the scholarship of teaching and learning by convening collaborators with diverse expertise and an expanded set of methods.

**For learning technology developers**, participation in the types of partnerships and collaborations described in this book can be used to expand their internal capacity for research and analytics as well as gathering new insights into issues surrounding the implementation of their products. Collaborating with researchers and practitioners can help technology developers gain a fuller understanding of the things that are important to teachers—their potential customers—and to supporting student learning. Moreover, if they design their learning system with the idea of being able to provide data collection and storage infrastructure that can be later analyzed efficiently, they will be better prepared to drive future enhancements of their products. We have found that a surprising number of learning technology products are developed and marketed widely without the capability to capture the kind of data that can inform teaching and learning. If developers understand how data captured by their technologies can be used to improve teaching and learning, they can be better equipped to collect and store data that can support continuous improvement of their products and how they are used.

### *Contents of This Book*

This book seeks to support both researchers, practitioners, and developers in applying data-intensive research methods to improve learning environments. Our goal is to offer scaffolds that a team can use to develop a research–practice partnership and use data with the rigor needed to make meaningful progress.

This introduction to our approach for merging data-intensive research, improvement science, and educational research will be followed by a description of the kinds of data that are available for use within research–practice partnerships in Chapter 2. Chapter 3 then introduces analytical techniques used in data-intensive research projects at an introductory level. In Chapter 4, we discuss issues of data privacy and security as well as approaches for using student data for research and improvement purposes. In Chapter 5, we describe the influences that have fostered an increased reliance on data and evidence in educational decision making and various conceptions of how researchers and education practitioners should work together in greater detail. These traditions influenced our own research and provided a foundation for our model of CDI. Using two cases, Chapter 6 presents our CDI model and discusses key assumptions of the model. Chapter 7 provides a deep dive into CDI practices and tools, presenting five phases for implementing this kind of work. Finally, we conclude in Chapter 8 with a summary of some of the key things we have learned from our work and the work of others and an explication of trends that are likely to shape future applications of data-intensive research in education.

## References

Agasisti, T., & Bowers, A. J. (in press). Data analytics and decision-making in education: Towards the educational data scientist as a key actor in schools and higher education institutions. In G. Johnes, J. Johnes, T. Agasisti, & L. López-Torres (Eds.), *Handbook on the economics of education*. Cheltenham, UK: Edward Elgar Publishing.

Allensworth, E. M., & Easton, J. Q. (2007). *What matters for staying on track and graduating in Chicago Public Schools*. Chicago, IL: University of Chicago Consortium on Chicago School Research.

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, *26*(2), 600–614. doi:10.1007/s40593-016-0105-0

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics* (pp. 61–75). New York: Springer. doi:10.1007/978-1-4614-3305-7_4

Balfanz, R., & Byrnes, V. (2012). *Chronic absenteeism: Summarizing what we know from nationally available data*. Baltimore: Johns Hopkins University Center for Social Organization of Schools.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. Washington, DC: U.S. Department of Education.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Press.

Coburn, C. E., & Penuel, W. R. (2016). Research-practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher*, *45*(1), 48–54.

Coburn, C. E., Penuel, W. R., & Geil, K. E. (2013). *Research-practice partnerships: A strategy for leveraging research for educational improvement in school districts*. New York: William T. Grant Foundation.

Dede, C. J. (Ed.). (2015). *Data-intensive research in education: Current work and next steps*. Computing Research Association. Retrieved from http://cra.org/wp-content/uploads/2015/10/CRAEducationReport2015.pdf

evoLLution. (2016). The scale of change in higher education: Using technology and impacting student success. Retrieved June 27, 2017, from https://evolllution.com/technology/tech-tools-and-resources/the-scale-of-change-in-higher-education-using-technology-and-impacting-student-success/

Ferenstein, G. (2016, January 20). *Why 'data scientist' is the best job to pursue in 2016*. Forbes. Retrieved October 1, 2016, from www.forbes.com/sites/gregoryferenstein/2016/01/20/report-why-data-scientist-is-the-best-job-to-pursue-in-2016/#7a1432f45f4b

Hanover Research. (2014). *Early alert systems in higher education*. Washington, DC: Author.

Hutchings, P., Taylor Huber, M., & Ciccone, A. (2011). *The scholarship of teaching and learning reconsidered: Institutional integration and impact*. San Francisco: Jossey-Bass.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–78). Cambridge: Cambridge University Press.

Koedinger, K. R., & McLaughlin, E. A. (2016). Closing the loop with quantitative cognitive task analysis. Presented at the *9th International Conference on Educational Data Mining*. Raleigh, NC. Retrieved from www.educationaldatamining.org/EDM2016/proceedings/paper_152.pdf

Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., & Nixon, T. (2013). Using data-driven discovery of better student models to improve student learning. In *Proceedings of the 16th international conference on artificial intelligence in education* (pp. 421–430). Phoenix, AZ.

Krumm, A. E., Waddington, R. J., Teasley, S. D., & Lonn, S. (2014). Using learning analytics to support academic advising in undergraduate engineering education.

In J. A. Larusson & B. White (Eds.). *Learning analytics: From research to practice* (pp. 103-119). New York: Springer.

Kurzweil, M., & Wu, D. D. (2015). *Building a pathway to student success at Georgia State University: A case study*. New York: ITHAKA S+R.

Martin, T., & Sherin, B. (2013). Learning analytics and computational techniques for detecting and evaluating patterns in learning: An introduction to the special issue. *Journal of the Learning Sciences*, *22*(4), 511–520.

O'Neil, C., & Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. Sebastopol, CA: O'Reilly Media.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students, vol. 2: A third decade of research*. San Francisco: Jossey-Bass.

Penuel, W. R., & Gallagher, D. J. (2017). *Creating research-practice partnerships in education*. Cambridge, MA: Harvard Education Press.

Piety, P. J., Hickey, D. T., & Bishop, M. J. (2014). Educational data sciences: Framing emergent practices for analytics of learning, organizations, and systems. In *Proceedings of the 4th international conference on learning analytics and knowledge* (Indianapolis, IN) (pp. 193–202). New York: ACM. doi:10.1145/2567574.2567582

Renick, T. (2017, May). *Keynote address for EdTech efficacy research academic symposium sponsored by University of Virginia Curry School of Education and Digital Promise*. Washington, DC.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146.

Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open*, *2*(4), 1–12. doi: 10.1177/2332858416673968

Shulman, L. S. (2000). From Minsk to Pinsk: Why a scholarship of teaching and learning? *The Journal of Scholarship of Teaching and Learning*, *1*(1), 48–53.

Siemens, G., & Baker, R. S. J. D. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In S. B. Shum, D. Gasevic, & R. Ferguson (Eds.), *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252–254). New York: ACM. doi:10.1145/2330601.2330661

To&Through Project. (no date). To&Through issue brief: Attendance. Retrieved June 28, 2017, from https://toandthrough.uchicago.edu/resources

Treaster, J. B. (2017, February 2). Will you graduate? Ask big data. *New York Times*. Retrieved from www.nytimes.com/2017/02/02/education/edlife/will-you-graduate-ask-big-data.html?_r=0

UT Arlington News Center. (2014, November). UT Arlington to lead $1.6 million research project focused on digital learning. Retrieved June 27, 2017, from www.uta.edu/news/releases/2014/11/LINKLab-dLRN.php

Chapter 2

# Data Used in Educational Data-Intensive Research

As early as the 1960s, computers began to fascinate educators. One of the first broadly implemented computer-based learning systems, PLATO (Programmed Logic for Automatic Teaching Operations), arrived 9 years before the first ARPANet transmission—the forerunner of the Internet—and 17 years before the Apple II popularized personal computing. As computers branched out beyond the realms of banking and scientific calculations and into personal applications, the idea of using computers to support teaching and learning gained widespread acceptance (Cuban, 1986). While interest was sparked early on, it took many years for technologies to become widely adopted and implemented with any depth in schools and universities (Collins & Halverson, 2009; Krumm, 2012). The story of technology integration in educational organizations intersects with data-intensive research in important ways: Some of the first technologies to be broadly adopted—learning management systems and intelligent tutoring systems—represent key touch points for the fields of learning analytics and educational data mining, respectively (Baker & Siemens, 2014).

Turning data into knowledge has until very recently been a manual activity (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Manual analysis has been the norm in schools and universities since the dawn of the Progressive era and the proliferation of Scientific Management practices (Tyack & Cuban, 1995). In more recent times, increased use of technology has led to the collection and storage of data that push on the capabilities of most manual approaches. In addition, as the volume of data has increased, so too has the need to combine data from across multiple technological platforms, like administrative data systems and digital learning environments, to better understand the processes and outcomes of teaching and learning. As combining data becomes more important, computer-based techniques are often required to merge, process, and analyze these data, all in an effort to unlock potential insights.

In this chapter, we introduce two of three foundational topics related to data-intensive research in education—*data* and *workflow*. In Chapter 3, we discuss the third foundational topic—analytical *methods*. None of these topics is unique to data-intensive research *per se*. For example, a workflow is a job-specific set of processes that transform inputs into outputs. All research involves some type of workflow—collecting data and analyzing it using recognized methods (i.e., inputs) to generate new knowledge (i.e., outputs). Over the next two chapters, we emphasize what is distinctive about data-intensive research across these general topics.

In what follows, we begin by outlining three general types of data used in educational data-intensive research. While we describe each somewhat in isolation, in practice, and in many of the examples that we cite, researchers regularly find value in combining different types of data. Following our discussion of three types of educational data, we discuss unique opportunities and challenges associated with using educational data as part of a data-intensive project. Based on our descriptions of three data types, we then introduce a generic workflow that outlines the ways in which data from multiple sources can be analyzed, interpreted, and translated into change ideas that are taken up as part of a formal research study or local improvement project.

## Types of Educational Data Used in Data-Intensive Research

In this section, we describe three broad types of data that are perhaps best characterized by the technologies in which they are captured and stored: (1) digital learning environments, (2) administrative data systems, and (3) sensors and recording devices. Data from *digital learning environments*, perhaps more than any other, have fueled data-intensive research in education (Roschelle & Krumm, 2015; Winne, 2017). Games, simulations, and tutoring systems as well as the increased amount of teaching and learning that is occurring through online courses and Massively Online Open Courses (MOOCs) are all creating more and more data on more and more students.

A second type of data fueling data-intensive research in education comes from *administrative data systems*. These systems are used in schools and districts as well as at the level of state and federal governments in the United States to collect and store information associated with delivering some type of service (Figlio, Karbownik, & Salvanes, 2017). For example, with investments from the U.S. Department of Education, states throughout the U.S. have created statewide longitudinal data systems that collect and store data on individual students over time. Data stored in these systems can include standardized test performances, attendance, and major behavioral infractions. Increasingly potent as tools for research, administrative data systems are creating opportunities for researchers and interested practitioners to jointly interpret data to both improve services as well as answer questions that are useful to the broader research community (Connelly, Playford, Gayles, & Dibben, 2016).

Lastly, as data from digital learning environments have been increasingly collected and stored, so too are data being collected from *sensors and recording devices*, such as video and audio data. Sensor and recording device data have increased in availability through newly developed instruments that capture biometric data and the ability to parse audio and video recordings using machine learning and artificial intelligence techniques. In education, data from sensors and recording devices have been combined with data from digital learning environments, like intelligent tutoring systems (e.g., Bosch, Chen, Baker, Shute, & D'Mello, 2015). These multiple data streams have been blended together to advance researchers' understanding of student learning and factors affecting learning over time within these environments.

## *Digital Learning Environments*

In the following sections, we highlight three digital learning environments based on the degree to which they are used in schools and universities and in their prominence in the research literature: intelligent tutoring systems, learning management systems, and MOOCs.

### *Intelligent Tutoring Systems*

An intelligent tutoring system (ITS) is a type of digital learning environment that applies artificial intelligence to students' interactions with the system. ITSs often employ three *models* that drive the adaptations that a system makes based on a student's input: (1) an expert, or *domain model*, which organizes the skills and strategies in the domain, (2) a *student model* of what a student understands about the domain that is inferred from their performances on learning tasks and (3) an *instructional*, or *pedagogical model*, of common mistakes and misconceptions along with a corresponding feedback strategy (Anderson, Corbett, Koedinger, & Pelletier, 1995). ITSs collect information on students, their progress in the system, and interactions that they engage in during a learning task. ITSs provide feedback to students in the form of hints, strategies, and different ways to practice the skills on which they need help (Razzaq & Heffernan, 2006). The same data that the ITS uses to figure out how to respond to a student's actions can also be used by human analysts to gain a detailed picture of learning processes and the behaviors learners engage in (Baker, 2016). For example, work by Baker and colleagues (Baker, D'Mello, Rodrigo, & Graesser, 2010) illustrate how data from ITSs can be used to detect a variety of behaviors and affective states such as boredom and frustration. These studies help in building knowledge related to how students learn as well as support potential improvements to the ITSs themselves (e.g., Roll et al., 2006).

### *Learning Management Systems*

LMSs are "web-based systems that allow instructors and students to share instructional materials, make class announcements, submit and return course assignments, and communicate with each other online" (Lonn & Teasley, 2009, p. 686). As noted previously, LMSs, along with ITSs, helped give rise to the fields of learning analytics and educational data mining, respectively. LMSs typically collect information on learning resources (e.g., digital files posted by an instructor) that students accessed and when they accessed them as well as when students accessed an assessment and how well they did on the assessment. Currently, LMSs are more widely used in higher education than in K–12, and they tend to be adopted on a campus-wide basis with the intent that all online and blended courses offered by a college or university are supported by the same LMS. Using data collected and stored by a campus's LMS, Krumm (2012) examined approximately 20,000 courses taught at the University of Michigan. Major takeaways from these analyses revealed that most instructors use relatively few tools that are provided by the LMS but that factors such as the college one teaches in and the enrollment size of the course can affect the number of tools used. In general, instructors favor using tools that make their teaching more efficient as opposed to rethinking how they teach (Lonn & Teasley, 2009). Said differently, while LMSs can be considered widely adopted, they are often not central to teaching and learning. However, when these systems are more central to instruction, researchers have found ways to use data from these systems to drive early warning systems. One such tool that allows instructors to provide feedback to students based on their interactions with an LMS is Course Signals, which was originally developed and deployed at Purdue University (see Arnold & Pistilli, 2012).

### *Massive Open Online Courses*

When MOOCs burst onto the higher education scene in 2010, course enrollments reached hundreds of thousands of students (Means, Bakia, & Murphy, 2014). Critics have been quick to point to the relatively small percentage of enrollees who actually completed these free online courses, and the hype around MOOCs has abated. Nevertheless, the MOOC learning platforms designed for very large enrollments, such as Coursera and edX, have endured, with large numbers of people taking courses on these platforms, including for academic credit. The data generated as thousands of learners use these platforms in a single course continues to be a major source of data for researchers (e.g., Evans, Baker, & Dee, 2016; Gasevic, Kovanovic, Joksimovic, & Siemens, 2014; Ho et al., 2015; Zhu et al., 2016). As these systems evolve, they continue to develop new features and functionality that capture granular data closer to ITSs than LMSs (e.g., Aleven et al., 2017).

## *Administrative Data Systems*

A second type of data fueling data-intensive research in education stems from *administrative data systems*. These systems are used at school and district levels as well as at the level of entire states. In this section, we describe two types of administrative data systems: student information systems and statewide-longitudinal data systems.

### *Student Information Systems*

Student information systems (SISs) are digital systems used by schools and universities to store student-level information. They are, in many ways, the central data repositories for educational organizations as they collect and store multiple data elements on students, including demographics, attendance, and academic performances. SISs are different from learning management systems, but the two can be integrated. While LMSs are often used as student-facing repositories of digital resources and activities, SISs are teacher- and administrator-facing repositories of student demographic and learning-outcome data. SISs play a key role in data-intensive research because they offer a ready source of data on educational outcomes (e.g., grades) and demographic information, which can play a role in evaluations of technology-based interventions as well as early warning system research (e.g., Bowers, Sprott, & Taff, 2013).

### *Statewide Longitudinal Data Systems*

In 2005, the U.S. Department of Education began giving grants to states to develop statewide longitudinal data systems (SLDSs). Among the requirements for SLDSs developed with these funds was the use of a unique statewide identifier for every student; storage of each student's demographic characteristics and enrollment history and scores on state accountability tests; and the ability to link the student's K–12 data with the state's higher education data system. According to the National Center for Education Statistics, by 2015, 84 percent of statewide longitudinal data systems contained unique student identifiers, 88 percent contained demographic and enrollment history data, and 57 percent could link to higher education data systems. For the first time, there was a data infrastructure in a majority of states that provided the potential to examine, for example, educational outcomes at the scale of an entire state. State level and university-based researchers increasingly leveraged these data for both accountability and reporting purposes as well as district-and school-improvement purposes. Knowles (2016), for example, used data from Wisconsin's SLDS to develop an early warning system for students at risk of dropping out of high school.

### *Sensors and Recording Devices*

As data from digital learning environments have been increasingly collected and stored, so too have data been collected from *sensors and recording devices*. Location, physical movement, and speech can all be tracked and analyzed using a variety of different sensors—small, often single data stream devices. Fitness sensors that measure, for example, steps taken or heart rate have been used in educational contexts to promote healthy behavior changes in youth (e.g., Schaefer, Ching, Breen, & German, 2016). Thus, sensor and recording device data have increased through instruments that capture biometric data, which are quantifications of an individual's physical activity. Moreover, the ability to parse audio and video recordings using machine learning and artificial intelligence techniques has opened up opportunities to analyze familiar forms of data, such as audio and video files, at larger and larger scales. Hand in hand with different algorithms have been multiple advances in collection and storage of these data in various digital formats (Baker & Siemens, 2014).

An important recent advance involves blending multiple data streams from sensors, recording devices, and digital learning environments (Blikstein, 2013; Liu, Davenport, & Stamper, 2010). Merging, or fusing, data from multiple systems can allow researchers to identify patterns across the different data streams that have been brought together. These *multimodal* investigations are providing new insights into basic factors affecting learning (e.g., Woolf et al., 2009). Understanding what learners do as they engage in learning tasks can drive digital learning environment adaptations. Recent work suggests that combining sensor data with data from digital learning environments can support accurately identifying multiple affective and engagement-related states (e.g., D'Mello, Dieterle, & Duckworth, 2017).

In the same way that data from sensors can be used to measure specific behaviors over time, audio and video recording data can be used by to detect facial expressions (e.g., Bosch, D'Mello, Ocumpaugh, Baker, & Shute, 2016) and body language (e.g., Grover et al., 2016). Audio data can be used for speech recognition, and even without analyzing the meaning of the recorded utterances, speech prosody (i.e., stress and intonation) can be used to make inferences about the emotional state of speakers.

For example, D'Angelo et al. (2015) are building speech-based learning analytics for collaboration that can support teachers to identify what is occurring in small groups, thereby enabling teachers to direct their attention to less well-functioning groups. Pilot data have shown that combining speech activity (i.e., who is talking when) with the actions of collaborators in digital learning environments can identify turn-sharing and frustration.

## Characteristics of Educational Data

The three types of education data described previously are intended to be overarching categories with which to think about the rapidly expanding types of data used to understand and improve teaching and learning. As can be seen in the previous examples, many researchers and research groups combine data from across these categories, and much of what are considered *big data* in education fall into one or more of the categories described previously. But what exactly are big data? Many scientific disciplines work with large, complex datasets (Dede, 2015), and the term big data is a relative and regularly shifting assessment of "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika et al., 2011). In a similar way, *data-intensive research* is a relative term that speaks to both the data and the research field in which the data are collected and analyzed. By current standards, datasets used in educational data-intensive projects have hundreds of thousands or millions of observations or hundreds or thousands of *features*. Depending upon the analytical method used, many of these datasets require software and hardware with specific capabilities to analyze, and the specific hardware and software will vary depending upon the ultimate purposes one has for an analysis.

In our own work, we regularly draw on data from LMSs and SISs. And over the years, we have developed a degree of familiarity with how to wrangle, explore, and model these types of data. Data from LMSs are often similar to one another but different from other types of educational data one could use in a data-intensive research project. Reasons for why these data are similar to one another but different from other types of data involve (1) the tasks that students are engaged in and (2) how data from those tasks are collected and stored by the technology. As noted previously, the types of data that are most often collected from LMSs include learning resources selected and when, as well as learning activities, such as assessments completed and when. These data can be substantively different from, for example, game-based learning environments because, at multiple levels, the types of activities that students are engaged in within a digital game are often dramatically different from an LMS (e.g., Owen, Ramirez, Salmon, & Halverson, 2014). Thus, working with and making sense of data require becoming familiar with the activities of the digital learning environment as well as the ways in which data from those activities are captured and stored for later use.

As one explores data from different technologies, one is likely to experience both structured and unstructured data—as well as variations in between. Structured data does not have a precise definition. In general, it is any kind of data organized into a table with rows and columns. Therefore, structured data have an explicit organization, and more often than not, structured data are housed in well-defined relational databases. One of the benefits of structured data is that they can more easily be manipulated and analyzed than unstructured data, such as large segments of text, audio, and video. While similarly lacking a precise definition, what makes unstructured data *unstructured* is that it does not have an explicit, predefined organization. Thus, tabular organization must be provided after the fact, often requiring significant wrangling and pre-processing. For example, when assessing samples of student writing, each sample needs to be converted into a list of numeric features, many thousands of them, each of which captures a different characteristic (Rutstein & Neikrasz, 2016). These numeric features can then be modeled using supervised machine learning algorithms across training and testing data. Known outcomes from individuals who scored the same writing samples train an algorithm. After adequate training and testing, the algorithm can be put into production in order to score new, unseen texts. Figure 2.1 illustrates this workflow. The latter part of this workflow is made possible by providing tabular, numeric structure to the previously unstructured data.

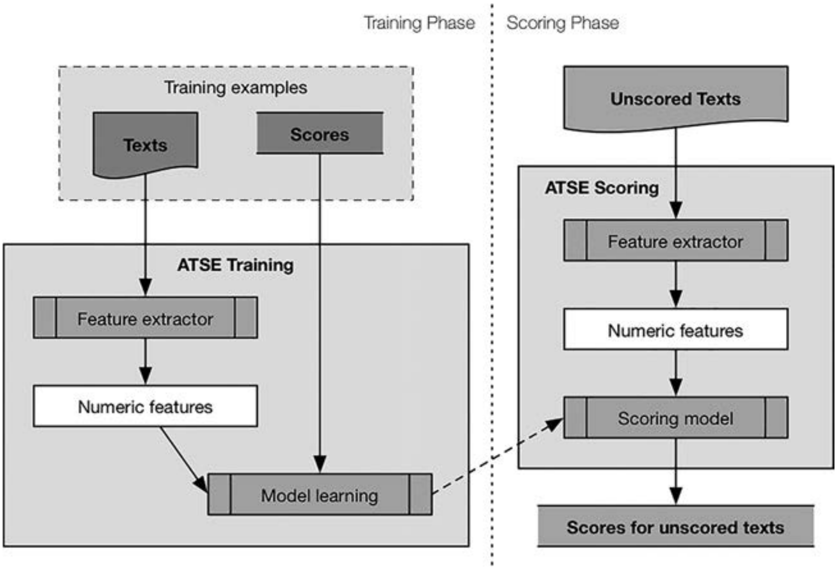


*Figure 2.1* Training and Scoring Phases in an Automated Text Scoring Engine (ATSE)

Oftentimes, discussions of big data in education mention anywhere from 3 to 7 “Vs”: volume, velocity, variety, veracity, variability, visualization, and value (e.g., van Rijmenam, 2013). The intent of these Vs is to help distinguish the types of data one is likely to use in data-intensive research as opposed to more traditional modes of inquiry. The four Vs that are most germane to data as opposed to how the data are used include volume, velocity, variety, and veracity. Volume is about the amount of data available, which is often affected by the number of observations, the number of features per observation, or both. Velocity addresses the rate at which data are generated; for example, every click that a student makes within a digital learning environment can lead to a rate of multiple clicks per minute, and over multiple minutes spread out over multiple days, volume and velocity can become closely related ideas. Variety describes the different types of observations, or events, that can be gleaned from a technology. For example, a digital game environment, from the same session, can continuously track a player’s screen coordinates as well as specific interactions within the game environment, all of which can lead to highly variable data over time. Veracity captures the degree to which a user can trust data. There are no standard units of measure for veracity, but data can be untrustworthy for a variety of reasons. For some administrative data systems, individuals inputting the data can use the system differently than intended, which means end-users of the data, such as researchers, need to understand as deeply as possible how and why data are entered into and stored within a system (Figlio et al., 2017).

To ground the four Vs in an educational example, as a researcher, imagine working with several high schools in a large district to help them in identifying patterns in students’ attendance over time. Instead of looking at whether a student was present or absent for the entire school day, the participating high schools are interested in the patterns that manifest by following individual students over time on a period-by-period basis. Teachers in participating high schools recorded whether or not a student was present across seven instructional periods, which yields seven measures per day, per student. For 6,000 students across the high schools, this would create a table with over 7.5 million cells for a 180-day school year. Following a single cohort of students from grades 9 through grade 12, then, would produce over 30 million unique observations that could be mined to look for patterns in which classes are missed most often and the relationships between missing class and overall school performance. Over a four-year period, the *volume* of a final dataset, depending upon when downloaded and analyzed, will reach the numbers identified previously. The *velocity* of these attendance data could best be thought of as hourly (i.e., at least during a typical school day). Importantly, these data are marginally low variety in the form of “present” or “not present” for a given hour of the day, such as “Period 1.” If one is interested in the specific courses a student was present or absent for, such as “Period 2 Geometry,” this level of detail increases the variety. *Veracity* is about the degree to which one can trust a row, column, or cell of data. For example: what does it mean to be counted as “late” for a class? The rule that defines lateness may or may not match the expectations of end-users of the data.

The four Vs described previously offer useful ways of thinking about characteristics of big data, but these characteristics may not ultimately address what can be unique about educational data used in a data-intensive research project. What is unique about working with educational data, especially within the context of formal schooling environments, is the interaction between the technology and the environment in which the technology is used. Focusing on the technology, we have noted the importance of the *tasks that students engage in* within a digital learning environment as well as the ways that *data from those tasks are collected and stored* by the technology. A rich digital learning task that does not capture granular data on what students do within the task, by definition, will not be useful for data-intensive research as requisite data are not collected. Less rich tasks that capture data on what students do, on the other hand, may also not be useful for data-intensive research as these data often fall victim to the garbage in, garbage out principle (Mislevy, Behrens, DiCerbo, & Levy, 2012). Rich tasks that are specifically developed so that students can generate meaningful events represent the best initial set of technology-specific circumstances for analyzing educational data (Schwartz & Arena, 2013; Shute & Ventura, 2013). However, technologies are not used in a vacuum—when taken up in schools, a technology will be used by students and teachers who can have different goals from those of the technology’s developer.

Two other characteristics of educational data based on the interaction between a technology and the environment in which it is used include *coverage* and *centrality*. Coverage as we use the term denotes the number of students within an educational organization who use a given technology from which data are collected. Centrality denotes the degree to which the technology is used as a core element, or facilitator, of instruction, i.e., how students interact with one another, the instructor, and content to be learned (Cohen, Raudenbush, & Ball, 2003). Coverage can be important because a technology from which data will later be used may in fact not be used by large numbers of instructors or students. Fewer students or more narrow groups of students (i.e., two dimensions of coverage) will dramatically affect the claims a researcher may seek to make based on the particular coverage of a technology he or she is analyzing. Given the diversity of content areas and instructional approaches in schools, the types of technologies with the most coverage by default tend to be those that facilitate more generic instructional interactions, such as accessing resources and submitting assessments. LMSs are a prototypical example of a broad coverage, generic technology as they can be used in all content areas and nearly all grades. Other technologies that offer high degrees of coverage include administrative data systems, as they track similar data elements for nearly all students.

Coverage and centrality, much like the four Vs described previously, are not inherently positive or negative characteristics. Broad coverage and non-central data from student information systems largely fuel early warning system research and development. Highly central technologies that have broad coverage are rare. One challenge in working with broad coverage systems like LMSs is that there are often large amounts of variation across educational units, such as classrooms and courses, using the technology. Given these differences, data from broad coverage technologies often necessitate that special attention be paid to unit-to-unit differences.

Task richness and how data are collected from tasks; coverage and centrality; the 4 Vs; and traditional considerations of quality educational research, such as overall research design, all factor into using educational data for data-intensive research. Building on these general characteristics, we now turn our attention to practical concerns around accessing and sharing educational data.

### *Challenges and Opportunities in Working With Educational Data*

There are multiple challenges as well as opportunities in working with educational data. Opportunities include building new knowledge as well as engaging in practical school improvement work—and new ideas yet to be developed. Data from digital learning environments as well as sensors and recording devices offer unique opportunities because they can be used to measure educational *processes* as they unfold over time. A core tenet of improvement is that changes in outcomes are dependent upon changes in processes (Langley et al., 2009). As many of the articles cited previously demonstrate, rigorous analyses of process data can also be used to build new understandings of how people learn. While there are a number of opportunities, privacy and security remain large and looming challenges in working with educational data. In Chapter 4, we detail many of these issues. For the purposes of understanding the types of data introduced previously, in this section, we describe several challenges and opportunities facing researchers in working with educational data as part of a data-intensive project.

A big challenge involves working with data from across multiple technologies, such as digital learning environments and administrative data systems. Issues of different identifiers used across technologies as well as duplicated entries can make merging datasets a labor-intensive and sometimes error-prone activity. A related challenge to this is making sure that the right students are present in the right datasets, which is often most noticeable after different datasets have been merged together. The integrity of samples of students directly implicates the veracity of educational data used in data-intensive research, as we described previously. The data a researcher eventually analyzes depends upon the business rules of the database as well as the informal rules around how individuals input and make use of data within these systems. For example, what counts as an "enrolled student" in a college course that uses an LMS can be far from clear-cut using LMS data alone. Thus, for projects geared toward predicting students who are likely to drop out of a course, corroborating data from across multiple sources can become a critical activity. Ultimately, it is where intended and actual uses for a technology conflict that working with data from across multiple datasets can prove problematic because intended uses are easy to communicate through data dictionaries and other written materials—informal and non-standard uses, less so.

Solutions to some of these challenges have included services offered by for-profit companies and industry groups that support normalizing student rosters across technologies (e.g., Clever and OneRoster). Other approaches include growing numbers of educational organizations developing and housing more and more data in data warehouses, which often contain common identifiers across databases. Moreover, there are growing standards movements that are intended to help create more common data models for administrative data systems (e.g., Ed-Fi Alliance) and digital learning environments (e.g., Experience API, Caliper). In general, these efforts address *interoperability*; programs such as the Schools Interoperability Framework (SIF) and Common Education Data Standards (CEDS) have emerged from consensus among vendors on how data can be exchanged across systems.

### *Accessing Data*

A challenge for both new and experienced researchers and educational data scientists is accessing data. State departments of education and individual school districts, starting in the early 2000s, began using administrative data systems and making their data available to researchers for well-defined research purposes. Once in the hands of researchers, these data were analyzed, reported on, and oftentimes destroyed in line with more or less well-defined data-use agreements. In many ways, data has been open to researchers with legitimate research purposes for a long time. Similarly, data collected by a digital learning environment could be accessed and analyzed by researchers with legitimate research purposes. In education and the physical and social sciences more generally, there is a growing movement where various datasets are being made publicly available. These efforts are moving data from servers once only accessible by researchers into public repositories that are creating opportunities for researchers to explore new questions and individuals new to data-intensive research in education to develop skills using often highly structured and well-documented datasets.

Under the labels of "open data" and "reproducible science," a variety of data sources are being opened up to broader audiences. The basic idea behind the open data movement is that anyone can access or use a dataset, and key to this movement is not just accessibility but usability. Making open data usable means making it accessible in machine-readable, structured, granular, and well-documented formats. On a case-by-case basis, individual research projects have made data available to external audiences (e.g., the Study of Instructional Improvement at the University of Michigan). These efforts can support replication of results as well as new explorations. Researchers in other sciences have proposed principles for enhancing the reproducibility of those results that are based on computational methods. They argue that while sharing data is useful, unless the computational software and workflow are also made available, the "computational reproducibility" of the findings cannot be assured. "Access to the computational steps taken to process data and generate findings is as important as access to data themselves" (Stodden et al., 2016, p. 1240).

The rise of structured, machine-readable data permits researchers to combine information or search for new patterns and new insights. The National Center for Education Statistics (NCES) is another resource for a variety of accessible, well-documented datasets (e.g., the Common Core of Data). More recently, datasets from tutoring systems and large online courses (e.g., MOOCs) are also being used in this way. For example, Harvard and MIT, in 2014, released de-identified data from open online courses, containing the original learning data from the 16 HarvardX and MITx courses offered in 2012–13 (Ho et al., 2015). Researchers at the Pittsburgh Science of Learning Center have developed and maintained the "world's largest repository of learning interaction data" in DataShop (Koedinger et al., 2010). DataShop contains data from multiple online educational environments, is open access, and is designed to provide researchers with a place to share data as well as analytic tools.

## Data-Intensive Research Workflow

The forerunner to data-intensive research, and therefore learning analytics and educational data mining, is a field of inquiry referred to as knowledge discovery in databases (KDD). The phrase was initially used in the late 1980s, and it was coined to emphasize that knowledge was the key outcome of any data-driven inquiry. From the outset, KDD referred to an overall workflow: "data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data" (Fayyad et al., 1996, p. 39). As we noted at the outset of the chapter, a workflow is a set of processes that transform inputs into outputs across multiple steps and decisions. A key input into this workflow consists of the types of data detailed previously. In this section, we introduce a generic workflow that is intended to support researchers, practitioners, and data scientists prepare for a data-intensive analysis and communicate one's findings. This workflow is based on workflows that have been documented by general data science practitioners (e.g., Guo, 2012; Wickham & Grolemund, 2017) as well as workflows that are based on practitioners' use of data in schools (e.g., Marsh, 2012).

A common workflow carried out using shared data analysis tools can make for efficient, reproducible data-intensive research (see Figure 2.2). In Chapters 6 and 7, we place this workflow within a broader set of phases that we use to help researchers and practitioners organize their collaboration around data-intensive analyses as well as co-developing and testing change ideas inspired by their analyses. The workflow described in the next sections comprises five steps: (1) prepare, (2) wrangle, (3) explore, (4) model, and (5) communicate. In Chapter 3, we go more in-depth into steps 2–4.

### *Prepare*

First and foremost, data-intensive research involves defining and refining one or more research questions. Having a clear set of research questions helps a team identify what data to collect and formulate potential analytical strategies. Along with clear questions, it can be useful to identify what gets collected and stored by a technology—not all potentially useful data are collected by a technology and not all data collected by a technology are useful. In an education context, understanding the *activity system* in

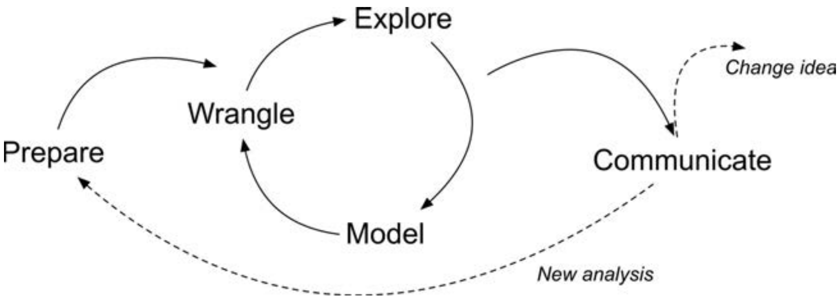


*Figure 2.2* Steps of Data-Intensive Research Workflow

which a technology is used can be crucial for ultimately making sense of data, in particular from digital learning environments (Roschelle, Knudsen, & Hegedus, 2009). Some instructional activity systems can include, among many factors, the actions and intentions of teachers and the goals that they have for students—from serving as a reward to students for completing work early to providing students' primary interactions with a course's content. All of these uses for a technology can affect the conclusions one can draw from data stemming from the technology as these different uses influence which students interact with it in the first place as well as what they do within the technology (Murphy et al., 2014). Being prepared for a data-intensive analysis, therefore, involves refining research questions and developing an understanding of where the data come from.

### *Wrangle*

Wrangling data, sometimes referred to as munging or pre-processing entails the work of *manipulating*, *cleaning*, *transforming*, and *merging* data. At a basic level, manipulating involves identifying, acquiring, and importing data into analysis software; cleaning data involves ensuring that each variable is in its own column, each observation is in its own row, and each value is in its own cell within a dataset (Wickham & Grolemund, 2017). Data cleaning also involves identifying and remediating missing data, extreme values, and ensuring consistent use of identifier, key, or linking variables. Data wrangling can also involve transforming variables, such as recoding categorical variables and rescaling continuous variables. These types of transformations are the initial building blocks for exploratory data analysis. Along with manipulation, cleaning, and transforming data, merging data is an important component of data wrangling. One of the earliest and biggest value-adds that a data scientist can bring to a formal research project or local improvement project is merging once disparate data sources. For example, merging data from a student information system that stores student grades with data from a digital learning environment that stores students' longitudinal interactions within a specific technology can be used to unlock the relationships between what students do or do not do on a day-to-day basis with how they performed on a longer-term outcome, such as a course grade. Merging data on what students do, i.e., process data, with how well they do, i.e., outcome data, are the building blocks of multiple types of *models*, described later.

### *Explore*

Exploratory data analysis is a widely covered topic that captures some combination of *data visualization* and *feature engineering*. Data visualization involves graphically representing one or more variables, whereby the goal of data visualization, according to Behrens (1997), "is to discover patterns in data that allow researchers to build rich mental models of the phenomenon being examined" (p. 154). Discovering patterns in data entails generating questions about one's data, visualizing relationships between and among variables, and creating as well as selecting features for subsequent data modeling. Feature engineering is the process of creating new variables within a dataset, which goes above and beyond the work of recoding and rescaling variables. For example, using data from an ITS, Baker, Gowda, and Corbett (2011) created new features, such as *the length of time a student paused after reading a hint*. Veeramachaneni, O'Reilly, and Taylor (2014) used brainstorming and crowd-sourcing techniques to develop features—such as *the difference in grade between current lab grade and average of student's past lab grade*—that were used to predict when students would stop actively participating in a MOOC course. Feature engineering draws on substantive knowledge from theory or practice, experience with a particular data system, and general experience in data-intensive research.

### *Model*

Modeling involves developing a mathematical summary of a dataset. There are two general types of modeling approaches: unsupervised and supervised learning. Unsupervised learning algorithms can be used to understand the structure of one's dataset. Supervised models, on the other hand, help to quantify relationships between features and a known outcome. Known outcomes are also commonly referred to as labels or dependent variables. A known outcome can include longer-term results of complex processes, such as dropping out of high school (Knowles, 2016), or shorter-term results like being off task (Hershkovitz, Baker, Gobert, Wixon, & Sao Pedro, 2013). Features used in a supervised learning model can also be referred to as predictors or regressors. Other names for features include attributes, independent variables, or simply—variables.

Unsupervised learning algorithms are often characterized as exploratory because unlike supervised learning models, they cannot be easily evaluated against a ground truth, or known outcome. When using supervised learning models, on the other hand, one can test a model's predictions against known outcomes. Supervised learning, or predictive modeling, involves two broad approaches: classification and regression. Classification algorithms model categorical outcomes (e.g., yes or no outcomes); regression algorithms characterize continuous outcomes (e.g., test scores). A model, the result of model-*ing*, can refer to either a general algorithm or a particular algorithm that has been applied to a particular dataset. When used to refer to a general algorithm, a model is a set of mathematical rules; in specific form, a model mathematically summarizes relationships within particular datasets (James, Witten, Hastie, & Tibshirani, 2013).

The process of modeling involves both *building* and *evaluation*. Building a model entails selecting features from a dataset and applying one or more algorithms to the dataset. Those who build a model are evaluating its performance using a variety of techniques, such as bootstrapping or cross-validation. Formally evaluating a model involves assessing its performance (i.e., how well it classifies categorical outcomes or predicts continuous values) on data that were not used to build the model. The steps involved in modeling, much like exploratory data analysis, are iterative and build on one another over time.

### Communicate

Communicating what one has learned involves *selecting* among those analyses that are most important and most useful to an intended audience. In addition, one must choose a form for displaying that information, such as a graph or table in static or interactive form. After creating initial versions of data products, research teams often spend time refining or *polishing* them, by adding or editing titles, labels, and notations and by working with colors and shapes to highlight key points. In addition, writing a *narrative* to accompany the data products is important and involves, at a minimum, pairing a data product with its related research question, describing how best to interpret the data product, and explaining the ways in which the data product helps answer the research question. These three steps—select, polish, and narrate—are intended to create a stand-alone data product that the intended audiences can use to inform their work.

The workflow cited previously lays out a series of steps for engaging in data-intensive research. Having a workflow creates multiple benefits and is intended to help both new and experienced educational data scientists create more reproducible data products, share analyses with internal and external audiences, and provide a structure for updating one's analyses over time. The workflow can help in achieving these goals by providing a key set of activities to address and an order in which to address them. While each step can and will be engaged in different ways across individuals and teams, each step represents an important one for almost any researcher or data scientist.

At the beginning of this section, we presented a somewhat linear movement across these five steps, from left to right in Figure 2.2. While there is often a great deal of iteration that occurs from wrangling to exploring to modeling, at any given time in a project one can be engaged in an activity that is difficult to put into any one step alone. Over time, we have come to see the workflow as overlapping activities as much as steps. Figure 2.3 is an alternative rendering of the workflow that captures the ways in which activities overlap and can be difficult to disentangle as

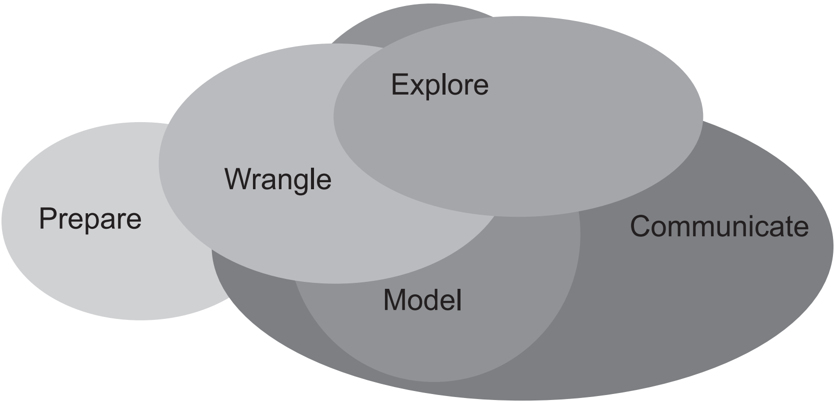


*Figure 2.3* Overlapping Activities Within the Data-Intensive Research Workflow

distinct steps—especially while engaged in a project. For example, *communicate*, in practice, is not a single step that occurs at the end of a formal modeling process. On the contrary, communication is happening throughout a project, and it is often only a matter of degrees that separates how much selecting, polishing, and narrating is involved in preparing for a research group's lab meeting and a formal presentation to a client or partner. Regardless of whether one is engaged in a formal research study or local improvement effort, when working with multiple complex datasets it is often the case that preparing, wrangling, exploring, modeling, and communicating will need to take place in more or less structured ways.

## Conclusion

The increasing use of technology in schools and universities is fueling the collection of ever more data on more and more students. Across learning environments of all kinds, there are three major sources of data that data-intensive researchers regularly draw upon: (1) digital learning environments, (2) administrative data systems, and (3) sensors and recording devices. In this chapter, we introduced a data-intensive research workflow that individuals and teams can draw on as they work with these types of data. This workflow is made up of five steps that address key elements of moving from identifying a dataset to producing a data product that answers an important question for researchers, practitioners, or both. This workflow will be used throughout this book. In the next chapter, we focus on three steps: wrangle, explore, and model and describe specific analytical techniques that researchers and data scientists can use in carrying out these steps.

## References

Aleven, V., Baker, R., Blomberg, N., Andres, J. M., Sewall, J., Wang, Y., & Popescu, O. (2017). Integrating MOOCs and intelligent tutoring systems: edX, GIFT, and CTAT. In *Proceedings of the GIFT Users Symposium (GIFTSym)*. Orlando, FL.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*(2), 167–207.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In S. B. Shum, S., D. Gasevic, and R. Ferguson (Eds.), *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267–270). New York: ACM.

Baker, R. S. J. D. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, *26*(2), 600–614.

Baker, R. S. J. D., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241.

Baker, R.S.J.D., Gowda, S.M., & Corbett, A.T. (2011). Automatically detecting a student's preparation for future learning: Help use is key. In *Proceedings of the 4th international conference on educational data mining* (pp. 179–188). Eindhoven, The Netherlands.

Baker, R. S. J. D., & Siemens, G. (2014). Educational data mining and learning analytics. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (2nd ed., pp. 253–274). Cambridge, UK: Cambridge University Press.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131.

Blikstein, P. (2013). Multimodal learning analytics. In *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (pp. 102–106) Leuven, Belgium, doi:10.1145/2460296.2460316

Bosch, N., Chen, H., Baker, R., Shute, V. J., & D'Mello, S. (2015). Accuracy vs. availability heuristic in multimodal affect detection in the wild. In *Proceedings of the 17th International Conference on Multimodal Interaction (ICMI)* (pp. 267–274). Seattle, WA.

Bosch, N., D'Mello, S.K., Ocumpaugh, J., Baker, R.S., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems*, 6(2), 1–26.

Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity and specificity. *The High School Journal*, *96*(2), 77–100.

Cohen, D., Raudenbush, S., & Ball, D. (2003). Resources, instruction, and research. *Education Evaluation and Policy Analysis*, *25*(2), 119–142.

Collins, A., & Halverson, R. (2009). *Rethinking education in the age of technology: The digital revolution and schooling in America*. New York: Teachers College Press.

Connelly, R., Playford, C. J., Gayles, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, *59*, 1–12.

Cuban, L. (1986). *Teachers and machines: The classroom use of technology since 1920*. New York: Teachers College Press.

D'Angelo, C., Roschelle, J., Bratt, H., Shriberg, E., Richey, C., Tsiartas, A., & Noyne, A. (2015). Using students' speech to characterize group collaboration quality. In O. Lindwall, P. Häkkinen, T. Koschman, P. Tchounikine, & S. Ludvigsen (Eds.), *Exploring the Material Conditions of Learning: The Computer Supported Collaborative Learning (CSCL) Conference 2015*, Volume 1 (pp. 819–820). Gothenburg, Sweden: The International Society of the Learning Sciences.

Dede, C. J. (Ed.). (2015). *Data-intensive research in education: Current work and next steps*. Computing Research Association. Retrieved from http://cra.org/wp-content/uploads/2015/10/CRAEducationReport2015.pdf

D'Mello, S. K., Dieterle, E., & Duckworth, A. (2017). Advanced, Analytic, Automated (AAA) measurement of engagement during learning. *Educational Psychologist*, *52*(2), 104–123.

Evans, B. J., Baker, R. B., & Dee, T. S. (2016). Persistence patterns in Massive Open Online Courses (MOOCs). *The Journal of Higher Education*, *87*(2), 206–242.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. *AI Magazine*, *17*(3), 1–30.

Figlio, D., Karbownik, K., & Salvanes, K. (2017). The promise of administrative data in education research. *Education Finance and Policy*, *12*(2), 129–136.

Gasevic, D., Kovanovic, V., Joksimovic, S., & Siemens, G. (2014). Where is research on massive open online courses headed? A data analysis of the MOOC research initiative. *The International Review of Research in Open and Distributed Learning*, *15*(5), 134–176.

Grover, S., Bienkowski, M., Tamrakar, A., Siddiquie, B., Salter, D., & Divakaran, A. (2016). Multimodal analytics to study collaborative problem solving in pair programming. In S. Dawson, H. Drachsler, & C. P. Rosé (Eds.), *Proceedings of the 6th international conference on learning analytics & knowledge* (pp. 516–517). New York: ACM.

Guo, P. (2012). *Software tools to facilitate research programming*. Unpublished Ph.D. dissertation manuscript, Stanford University, Stanford, CA.

Hershkovitz, A., Baker, R. S. J. D., Gobert, J., Wixon, M., & Sao Pedro, M. (2013). Discovery with models: A case study on carelessness in computer-based science inquiry. *American Behavioral Scientist*, *57*(10), 1480–1499.

Ho, A. D., Chuang, I., Reich, J., Coleman, C. A., Whitehill, J., Northcutt, C. G., . . . Petersen, R. (2015). HarvardX and MITx: Two years of open online courses fall 2012-summer 2014. *SSRN Electronic Journal*. doi:10.2139/ssrn.2586847

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (1st ed.). New York, NY: Springer.

Knowles, J. E. (2016). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining*, *7*(3), 18–67.

Koedinger, K. R., Baker, R. S. J. D., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. D. Baker (Eds.), *Handbook of educational data mining* (pp. 43–55). Boca Raton, FL: CRC Press.

Krumm, A. E. (2012). *An examination of the diffusion and implementation of learning management systems in higher education*. Unpublished Ph.D. dissertation manuscript, The University of Michigan, Ann Arbor, MI.

Langley, G. J., Moen, R., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance*. New York, NY: Jossey-Bass.

Liu, R., Davenport, J., & Stamper, J. (2016). Beyond Log Files: Using Multi-Modal Data Streams Towards Data-Driven KC Model Improvement. In *The 9th International Conference on Educational Data Mining (EDM 2016)* (pp. 436–441). Raleigh, NC.

Lonn, S., & Teasley, S. (2009). Saving time or innovating practice: Investigating perceptions and uses of learning management systems. *Computers Education*, *53*, 686–694.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey & Company. Retrieved from www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, *114*(11), 1–48.

Means, B., Bakia, M., & Murphy, R. (2014). *Learning online: What research tells us about whether, when, and how*. London and New York: Routledge.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., & Levy, R. (2012). Design and discover in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, *4*(1), 11–48.

Murphy, R., Snow, E., Mislevy, J., Gallagher, L., Krumm, A. E., & Wei, X. (2014). *Blended learning report*. Menlo Park, CA: SRI Education.

Owen, V. E., Ramirez, D., Salmon, A., & Halverson, R. (2014). Capturing learner trajectories in educational games through ADAGE (Assessment Data Aggregator for Game Environments): A click-stream data framework for assessment of learning in play. Presented at the *2014 American Educational Research Association Annual Meeting*. Philadelphia, PA.

Razzaq, L., & Heffernan, N. T. (2006). Scaffolding vs. hints in the Assistment system. In M. Ikeda, K. Ashley, & T. Chan (Eds.), *Proceedings of the eight international conference on intelligent tutoring systems* (pp. 635–644). Berlin: Springer-Verlag.

Roll, I., Aleven, V., McLaren, B. M., Ryu, E., Baker, R. S. J. D., & Koedinger, K. R. (2006). The help tutor: Does metacognitive feedback improve students' help-seeking actions, skills, and learning? In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 360–369). Jhongli, Taiwan.

Roschelle, J. & Krumm, A. E. (2015). Infrastructures for improving learning in information-rich classrooms. In P. Reimann, S. Bull, M. Kickmeier-Rust, R. Vatrapu, & B. Wasson (Eds.), *Measuring and visualizing learning in the information-rich classroom* (pp. 3–9). New York: Routledge.

Roschelle, J., Knudsen, J., & Hegedus, S. (2009). From new technological infrastructures to curricular activity systems: Advanced designs for teaching and learning. In M. J. Jacobson & P. Reimann (Eds.), *Designs for learning* *environments of the future: International perspectives from the learning sciences* (pp. 233–262). New York: Springer.

Rutstein, D. W., & Niekrasz, J. (2016). Automated scoring of constructed response items measuring computational thinking. Presentation at the *National Council on Measurement in Education*, Washington, DC.

Schaefer, S. E., Ching, C. C., Breen, H., & German, B. J. (2016). Wearing, thinking, and moving: Testing the feasibility of fitness tracking with urban Youth. *American Journal of Health Education*, *47*(1), 8–16.

Schwartz, D. L., & Arena, D. (2013). *Measuring what matters most: Choice-based assessments for the digital age*. Cambridge, MA: MIT Press.

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: MIT Press.

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., et al. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240–1241.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.

van Rijmenam, M. (2013, August 6). *Why the 3V's are not sufficient to describe big data* [Weblog post]. Retrieved from https://datafloq.com/read/3vs-sufficient-describe-big-data/166

Veeramachaneni, K. O'Reilly, U.-M., Taylor, C. (2014). *Towards feature engineering at scale for data from massive open online courses*. arXiv preprint arXiv:1407.5238

Wickham, H., & Grolemund, G. (2017). *R for data science*. Sebastopol, CA: O'Reilly Media.

Winne, P. H. (2017). Leveraging big data to help each learner and accelerate learning science. *Teachers College Record*, *119*, 1–24.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: Recognising and responding to student affect. *International Journal of Learning Technology, 4*(3/4), 129–164.

Zhu, M., Bergner, Y., Zhang, Y., Baker, R., Wang, Y., & Paquette, L. (2016). Longitudinal engagement, performance, and social connectivity: A MOOC case study using exponential random graph models. In S. Dawson, H. Drachsler, & C. P. Rosé (Eds.), *Proceedings of the 6th international conference on learning analytics & knowledge* (pp. 223–230). New York: ACM.

Chapter 3

# Methods Used in Educational Data-Intensive Research

Data from digital learning environments, administrative data systems, and sensors and recording devices have all helped to fuel the growth of data-intensive research in education. Yet, data are not the only contributors to the expanding fields of learning analytics and educational data mining—a rapidly expanding set of analytical methods are also supporting the growth of data-intensive research in education (Madhavan & Richey, 2016). In this chapter, we build on the workflow introduced in Chapter 2 and describe the steps of wrangle, explore, and model in greater detail. When paired with a more explicit how-to guide for software such as R (Wickham & Grolemund, 2017), Weka (e.g., Witten, Frank, Hall, & Pal, 2017), Python (VanderPlas, 2017), or RapidMiner (Kotu & Deshpande, 2014), to name but a few popular data analysis tools, this chapter is intended to help researchers organize their use of various analytical methods within an overall data-intensive research project and to further elaborate on the intuition behind wrangling, exploring, and modeling data.

Throughout our discussion of various methods, we use a hypothetical case to illustrate what can be involved at each step. As researchers, we regularly work with data from digital learning environments under a particular arrangement referred to as a research-practice partnership (RPP, e.g., Penuel & Gallagher, 2017). In most of the RPPs in which we work, we collaborate with educational organizations to help them learn from their own data and identify new ways to support students. Typically, our goal is not to engage in basic research—i.e., theory building—but to engage in more applied research—i.e., problem solving—that is informed by and in some cases based almost entirely on the needs of practitioners. In Chapter 5, we provide a more detailed account of the growth in RPPs as a strategy for engaging in educational research, and in Chapters 6 and 7, we go into more depth on how to launch and sustain an RPP organized around a data-intensive research project.

## Hypothetical Case

Suppose that, as researchers, we are brought together with leaders of a large urban district that is piloting a dual-enrollment math course where students can earn both high school and postsecondary credit. In keeping with the workflow introduced in the previous chapter, we prepare for this project by, first, clarifying a purpose and driving questions for the project and, second, understanding the digital learning environment and how it is used to support teaching and learning in classrooms. One high school in the district is participating in the program, and in its second year, the goal is to reduce the number of students who earn a C– or lower in the course.

For this small-scale project, we will analyze data from the first year of the dual-enrollment course, which served 100 students. Earning a C or higher is an important outcome to track and improve upon because a grade of C– or lower means that the student cannot count the course toward his or her postsecondary degree. Our role in this project is to analyze data that were collected and stored in a learning management system (LMS) that was used to deliver aspects of the math course. The assumption of practitioners in the high school is that the data generated on a day-to-day basis, or lack thereof, by students might reveal certain patterns that are indicative of students who are not likely to earn a C or higher. If these patterns could be identified early in the year, then a teacher could work with students sooner and hopefully prevent them from earning less than a C in the course.

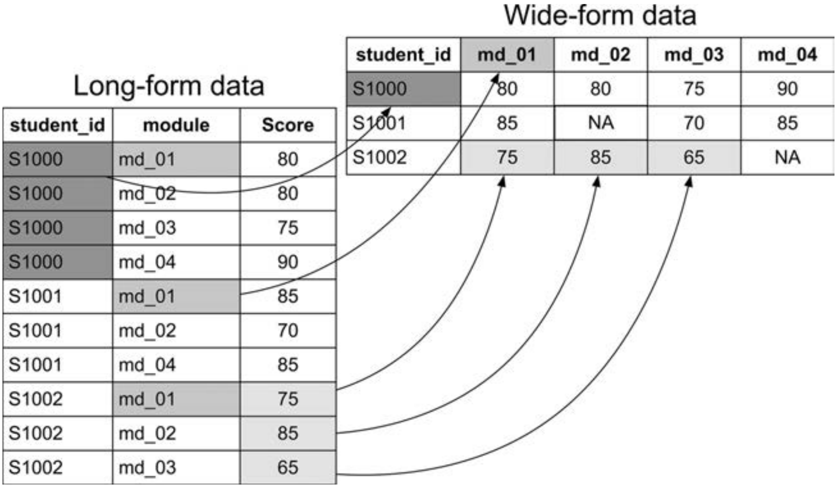
Course content in the LMS is organized into "modules" that comprise digital readings, practice activities, and a summative assessment. Each module covers a specific math topic and each student has full discretion to work on a module at his or her own pace. Students need to complete all nine summative assessments in the course. Performance on the assessments, other course assignments, and class attendance all contribute to a student's grade. Students can take the summative assessment multiple times, as each summative assessment attempt selects items from a semi-random item bank. For the first year of the dual-enrollment program, the LMS only accurately collected and stored data on students' summative assessment taking, which is what we can use, along with students' final grades, to understand factors affecting students' performances in the course and to eventually develop a predictive model that will help teachers identify students earlier in the year.

## Wrangle

At a general level, data wrangling involves some combination of cleaning, reshaping, transforming, and merging data (Wickham & Grolemund, 2017). Data wrangling can require knowledge of databases as well as data analysis software languages like SQL, R, and Python. In certain projects, data wrangling skills may be distributed across multiple individuals, such as those well versed in SQL and others in software like R. A component of engaging in data-intensive research is that oftentimes no one person holds all of the knowledge necessary for conducting an analysis (Piety, Hickey, & Bishop, 2014). For many projects, the importance of data wrangling is difficult to overstate, as it involves the initial steps of going from raw data that can be contained in multiple, distributed tables to a dataset that can be explored and modeled. Moreover, an individual's, or a team's, wrangling skills can make later exploration and modeling more efficient. Effective and efficient data wrangling is often programmed or scripted, which means that the steps involved in importing, cleaning, and merging data are written out as machine-readable commands that can be revisited, repurposed, and debugged over time. Data wrangling often runs throughout a data-intensive research project. For example, it is critical for feature engineering, which is described later on in this chapter. In later stages of a project, wrangling can entail extracting, cleaning, and merging data products. For example, being able to extract and organize output that is the result of, in some cases, the running of hundreds of models can aid in communicating findings later on in a project.

### *Clean*

Rarely are data accessed from an external source received in a form that is ready for analysis. Even structured data from well-maintained relational databases require cleaning, which involves identifying missing data as well as extreme or unexpected values. Thus, after extracting data from a source and importing it into a piece of software, cleaning is typically the first step in getting to know one's data. Getting to know one's data also involves determining how a data file is structured in either *long* or *wide* formats. Using students' interactions in an online learning system as an example, long-form data include multiple observations for one student across multiple rows. Figure 3.1, for example, illustrates four unique observations for student **S1000**. Each row contains a student identifier, a module identifier, and a score for a summative assessment. Long-form data typically have fewer columns and more rows than wide-form data, which is represented on the right-hand side of Figure 3.1. For wide-form data, each student occupies one and only one row. In the wide-form table, individual module assessments (e.g., **md_01**) make up the remaining columns, which contain scores in each cell for a given student. Moving from long- to wide-form data can go by various names that can be software specific, but in general, moving back and forth between long- and wide-form data is referred to as reshaping one's data.



Wide-form data

| student_id | md_01 | md_02 | md_03 | md_04 |
|---|---|---|---|---|
| S1000 | 80 | 80 | 75 | 90 |
| S1001 | 85 | NA | 70 | 85 |
| S1002 | 75 | 85 | 65 | NA |

Long-form data

| student_id | module | Score |
|---|---|---|
| S1000 | md_01 | 80 |
| S1000 | md_02 | 80 |
| S1000 | md_03 | 75 |
| S1000 | md_04 | 90 |
| S1001 | md_01 | 85 |
| S1001 | md_02 | 70 |
| S1001 | md_04 | 85 |
| S1002 | md_01 | 75 |
| S1002 | md_02 | 85 |
| S1002 | md_03 | 65 |

*Figure 3.1* Transition from Long- to Wide-Form Data

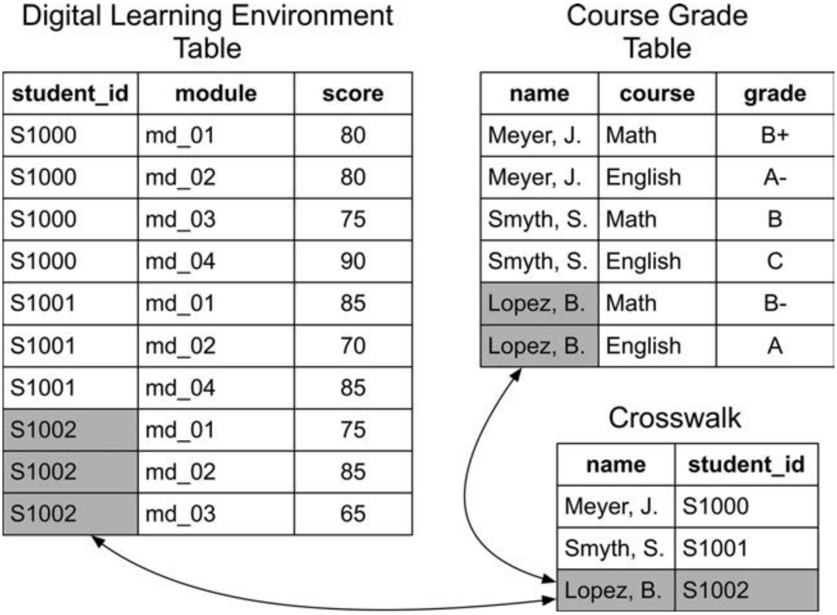
As can be seen in Figure 3.1, cleaning and reshaping data can reveal missing data, which is most noticeable in the wide-form version of the data (i.e., **NA**). Student **S1001**, for example, is missing assessment **md_02**. Missing data has long been a topic of discussion, especially in statistics (Rubin, 1976). Reasons as to why data are missing are key to solving potential problems brought on by missing data. When data are missing, as with **S1001**, the entire observation can be dropped from a given analysis, which can be less than advantageous because dropping observations can affect the representativeness of a sample. For students in our case example, not completing assignments is important to understanding overall course performance, so we do not want to drop students who miss assignments from our analyses. Thus, a key topic for many statistical and machine learning analyses is *imputation*, which involves substituting a missing value with another value, allowing the researcher to keep the overall observation in an analysis. For example, in Figure 3.1, one could impute **S1001**'s score for module assessment **md_02** by replacing the "NA" value with the overall average score for **md_02** and myriad other approaches (e.g., van Buuren & Groothuis-Oudshoorn, 2011). While there are many imputation approaches one can use, for the purposes of the above example, replacing "NA" with "0" allows for keeping the entire observation while providing a contextually relevant value. Each imputation approach, from the simple to complex, involves tradeoffs of one type or another and can be more or less appropriate based on the underlying reasons as to why data are missing.

Data exploration is often a necessary part of data cleaning. When done at this early stage, as opposed to the more formal exploratory data analysis step described later, data exploration is typically done to answer a basic question, *Does this file meet my expectations?* For example, one should expect consistency in the types of values within a given feature, such as all values being numeric. Along with common formatting, one may expect that values within a feature all fall within the bounds of an expected range, and if data are categorical, that values are all within predefined categories. Lastly, data may not meet one's expectations based on the number of observations present in a dataset. Peng (2016) recommends, whenever possible, finding a way to corroborate what is in a dataset with outside information. When working with data from schools, this regularly involves corroborating the number of students in a dataset with, for example, available class rosters and having follow-up conversations with teachers and administrators.

### *Merge*

Bringing together, or joining, data from different sources is a critical component of most data-intensive research projects. Many data quality issues can be identified—or introduced—as one merges data from multiple sources. For example, merging can bring to light observations that are missing in one dataset but present in another. Merging datasets requires common identifiers, often referred to as key variables, across the multiple datasets that one is trying to merge. Sometimes different identifiers for the same individual are used across datasets. For example, a student's email address may be in one dataset and his or her name may be used in another. In such cases, researchers need to create a table, or crosswalk, that aligns all of the identifiers for a given individual. Multiple, related issues can surface as one seeks to merge data from multiple systems, and for many projects, de-identifying a dataset (i.e., removing personally identifiable information) is another step that needs to be taken prior to merging files. Figure 3.2 illustrates how data stored in a **Course Grade Table** from a student information system can be merged with the **Digital Learning Environment Table** using a **Crosswalk** that aligns a student's **name** from the **Course Grade Table** to the **student_id** used in the digital learning environment.

Merging data can create the opportunity to surface insights that were not possible by looking only at any one dataset in isolation. For example, data from digital environments can contain information on what students do on a day-in and day-out basis, and data from an administrative data system can contain information on valued outcomes, such as students' final course grades. Feng, Heffernan, and Koedinger (2009), for example, merged data from the ASSISTments system and statewide-standardized assessments in Massachusetts to explore the benefits of using continuous assessment systems like ASSISTments over and above once-per-year



Digital Learning Environment Table

| student_id | module | score |
|---|---|---|
| S1000 | md_01 | 80 |
| S1000 | md_02 | 80 |
| S1000 | md_03 | 75 |
| S1000 | md_04 | 90 |
| S1001 | md_01 | 85 |
| S1001 | md_02 | 70 |
| S1001 | md_04 | 85 |
| S1002 | md_01 | 75 |
| S1002 | md_02 | 85 |
| S1002 | md_03 | 65 |

Course Grade Table

| name | course | grade |
|---|---|---|
| Meyer, J. | Math | B+ |
| Meyer, J. | English | A- |
| Smyth, S. | Math | B |
| Smyth, S. | English | C |
| Lopez, B. | Math | B- |
| Lopez, B. | English | A |

Crosswalk

| name | student_id |
|---|---|
| Meyer, J. | S1000 |
| Smyth, S. | S1001 |
| Lopez, B. | S1002 |

*Figure 3.2* Multiple Data Tables and Crosswalk

standardized assessments. In the context of evaluating a variety of digital learning environments, Murphy et al. (2014) combined data from multiple digital learning environments with data from students' standardized test performances. The growing use of predictive modeling and use of early warning systems regularly require researchers to merge datasets from multiple systems (e.g., Knowles, 2015). Moreover, the similarly growing use of predictive models to develop "behavior detectors" often requires researchers to merge data from digital learning environments and sensors and recording devices (e.g., Paquette et al., 2015) with data from observations taken in classrooms (e.g., Baker, Corbett, & Koedinger, 2004) or coded replays of students' use of a technology (e.g., Sao Pedro, Baker, Gobert, Montalvo, & Nakama, 2011).

## Explore

Exploratory data analysis often involves some combination of *data visualization* and *feature engineering*. These can be considered exploratory activities because they are often not done to formally test or confirm a hypothesis, but instead to help one continue to develop a richer understanding of one's data. When one first begins working with a dataset, especially after merging once disparate tables, it can be useful to begin visually exploring variation within individual variables and co-variation between two or more variables (Loeb, Dynarski, McFarland, Morris, & Reardon, 2017). Exploratory data analysis is a widely used phrase that can evoke specific steps and decisions to follow, or at a minimum, general rules of thumb. For example, Behrens (1997), building off of the work of Tukey (1977), outlines a somewhat formal approach for engaging in exploratory data analysis that is built on (1) understanding the context from which data were collected, (2) using graphical tools to understand the structure of and relationships within a dataset, (3) manipulating and creating features, and (4) iterative model building and evaluation.

Given the ways in which most discussions of exploratory data analysis draw on the work of John Tukey, many of the processes and styles of thinking that he introduced are present in recent discussions of exploratory data analysis. For example, Wickham and Grolemund (2017) describe exploratory data analysis as (1) transforming data (i.e., manipulating and creating features), (2) using data visualization tools, and (3) engaging in iterative model building. For our purposes, we draw particular attention to the ways in which visualization tools and feature engineering, combined, are central to exploring one's data. As we noted in Chapter 2, there are many overlaps across the steps involved in the data-intensive workflow. Feature engineering, for example, is closely related to data wrangling. Our primary reason for pairing feature engineering with data visualization—under the label of exploration—is that as one engages in creating new features, visualizing relationships among features can provide tentative evidence for a feature's value and can also help in identifying new features.

In exploring LMS from our participating high school, we want to identify the proportion of students who earned a C– or lower. To visualize this group of students, we can recode all of the different letter grades that students earned into a simple indicator (C– or lower = 0; C or higher = 1) and visually explore the number of students in each category. As demonstrated in Figure 3.3, 25 out of 100 students did not earn a C or higher.

To start engineering features related to students' success in the course, we can talk with teachers who taught the course last year to get a sense of what they perceived as contributing to the high numbers of students doing less well in the course. One potential contributing factor is that students started slow and never caught up. To explore this hypothesis, we can create a feature that addresses when a student first scored at or above 60 percent on a summative assessment. Using this feature and boxplots, we can observe when students, across modules, attained a minimal level of proficiency on a respective summative assessment (see Figure 3.4). This initial visualization draws attention to the somewhat linear pattern for when students first passed a summative assessment. Boxplots represent multiple points in a distribution created by a group of observations. The box represents the 75th and 25th percentiles and the line within the box represents the median, or 50th percentile. The lines, or whiskers, extend

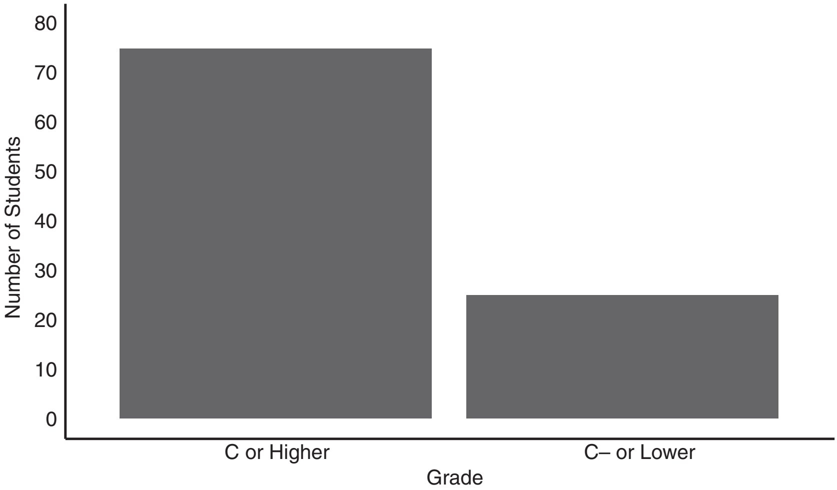


*Figure 3.3* Bar Chart of Earning a C or Higher

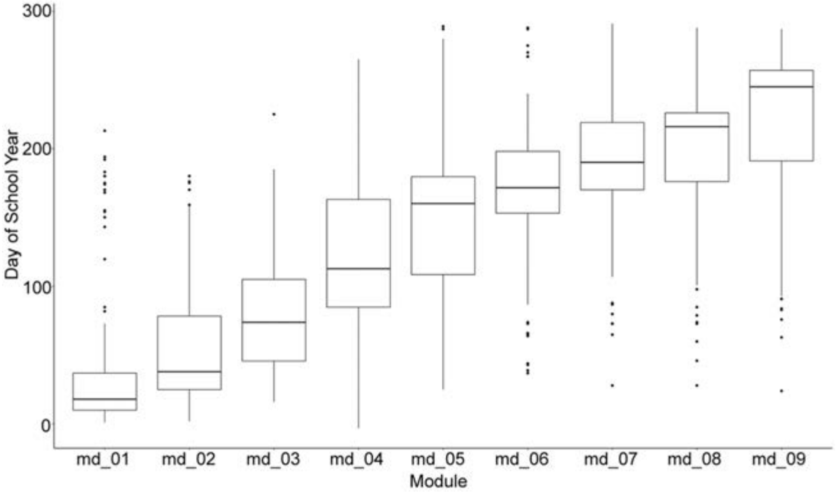


*Figure 3.4* Day of School Year for First Passing Summative Assessment

out from the 75th and 25th percentiles to represent the last non-outlier data points in either direction, which are defined as greater than 1.5 times the interquartile range (i.e., the difference between the 75th and 25th percentiles). Some modules, such as **md_04**, illustrate how students passed the summative assessment at almost all points throughout the year, from the very first days of school (i.e., close to 0) to the very last days of school (i.e., close to 300).

### *Engineer Features*

Feature engineering is a highly important but not often talked about element of engaging in data-intensive research (Paquette, de Carvalho, Baker, & Ocumpaugh, 2014). As a process with inputs, steps, and outputs, it involves using theory, knowledge from practice, and logic in order to create new features from existing datasets. These features can be used in a variety of data products, from predictive models and visualizations to tables comprising descriptive statistics. Thus, newly developed features need not only be used in a predictive model. For example, when combined with unsupervised learning approaches, new features can reveal important new groups within one's data. In many ways, feature engineering is where data wrangling and data exploration intersect. As a process, feature engineering is related to but distinct from *knowledge engineering*, which involves using theory and approaches, such as cognitive task analysis, to develop a representation of how to execute a process *a priori* (e.g., Paquette, de Carvalho, Baker, & Ocumpaugh, 2014). The term "knowledge engineering" is often used to refer to connecting different features together around complex constructs like "help seeking" (e.g., Aleven, Roll, McLaren, & Koedinger, 2016).

Feature engineering can be constrained by the complexity and granularity of available data. Some systems, like intelligent tutoring systems, provide granular depictions of what students do across a large number of actions as well as the system's response to a student's action. Some LMSs can achieve this level of granularity, but more often than not LMSs track events such as accessing a learning resource or when a learner has completed an assessment—as opposed to the items selected and hints used. Important to feature engineering, however, is not the volume of data, alone. As we described in Chapter 2, using data from digital learning environments is dependent upon the richness of the underlying task and the degree to which that richness is represented in the data that are collected and stored by the system. For example, there can be some digital environments, such as games, that produce large volumes of data on moment-to-moment position changes on the screen. However, unless this flow of data is in some way understandable in relation to the tasks in which learners are engaged and is directed at measuring a meaningful construct, it can be difficult to first operationalize these large volumes of data into features. For example, DiCerbo (2014) and Ventura, Shute, and Small (2014), in separate digital learning environments, measured "persistence" in ways that advanced thinking around both learning and assessment because of the quality of the underlying tasks and the ways in which data from those tasks were captured by their respective environments.

### *Data Visualization*

Visualization is key to both understanding one's data and communicating what has been learned. Tufte (2001), for example, identified and consolidated effective guidelines for developing quality data visualizations, such as having high "data-ink" ratios, highlighting comparisons, clarifying potential mechanisms operating in one's data, and illustrating multivariate relationships. Through cycles of feature engineering and data visualization, clearer comparisons, mechanisms, and relationships can manifest all while working to include as much raw data as possible (i.e., high data-ink ratio). For some projects, the focal end product is a visualization; a quality visualization that highlights a key disparity or opportunity for improvement can be an important catalyst for action. Thus, after exploring data, one can take steps toward communicating to a specific audience. One such example comes from a project where researchers were investigating ways that data-intensive research methods could be used to enhance stakeholders' decision making related to students' success in schools. Krumm, Boyce, Gassman-Pines, Bellows, and Podkul (2017) used a Sankey, or flow, diagram to illustrate the degree to which students in grades 6–8 from different economic backgrounds were formally written up for behavioral infractions in North Carolina schools. This diagram was built using administrative data from two separate state-level departments, the Department of Health and Human Services and the Department of Public Instruction. Students' participation in the Supplemental Nutrition Assistance Program (SNAP, or "food stamps") was identified using Department of Health and Human Services data, and students' behavioral infraction data came from the Department of Public Instruction. Sankey diagrams illustrate the movement of inputs across key steps, changes, or decision points making up a flow of activity whereby the width of various flows is proportional to the quantity of inputs. In the case of Figure 3.5, inputs are students, and the flows connect changes in SNAP status to being formally written up for an infraction. There are multiple ways to interrogate the visualizations, which makes them powerful tools when working within a partnership arrangement as was the case for the research team working with practitioners in the Department of Public Instruction. For example, even though black students made up only 26.3 percent of the overall student population in grades 6–8 in the 2011–12 school year, these students made up 39.1 percent of the students who were formally written up for a behavioral infraction in school. White students on the other hand, made up 53.7 percent of the overall population and only 42 percent of students who were formally written up. These observations led to follow-up analyses directed at understanding these discrepancies as well as efforts to better understand the relationships between being written up and academic performance.

Returning to our case high school, we have created and visualized a feature for when students first scored over 60 percent on a summative assessment. In continuing to visually explore the importance of students falling behind, we can compare the day of the year students first passed an assessment using the grade they eventually earned in the course. Thus, we can expand upon Figure 3.5 by creating two boxplots for each module.

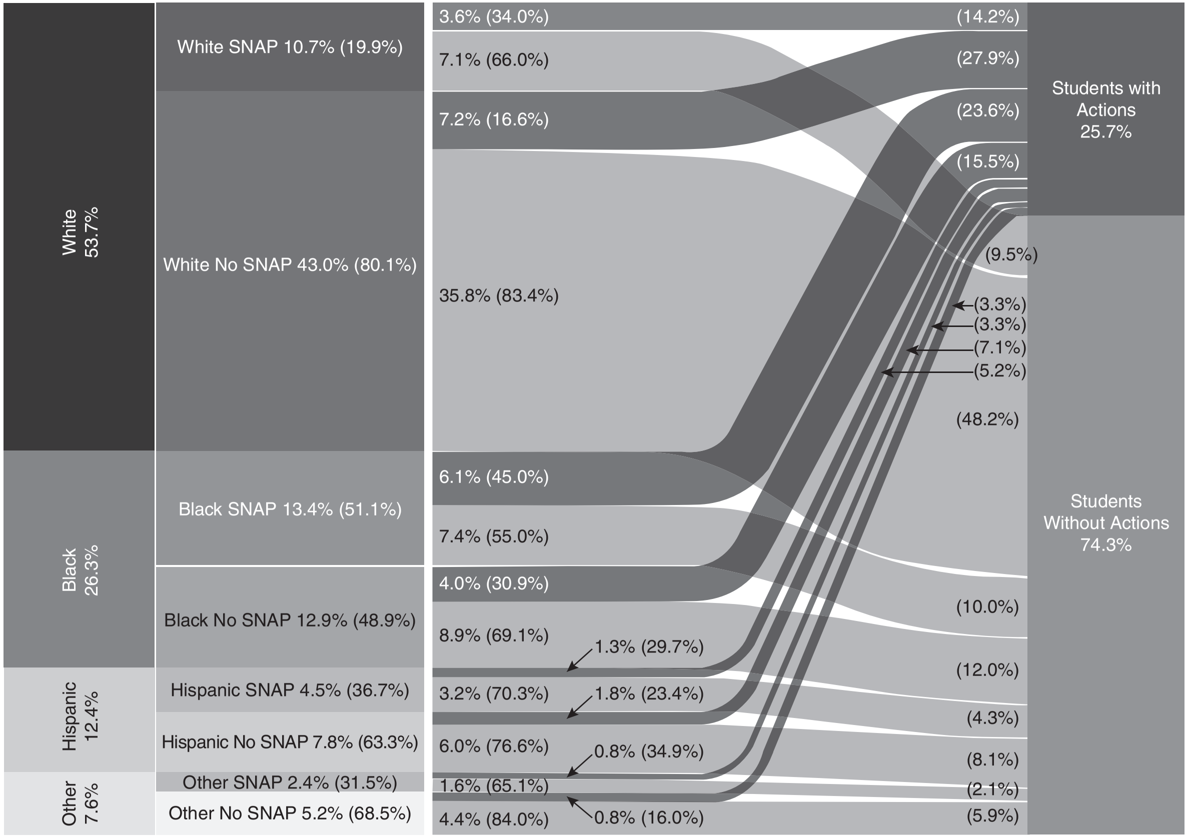


*Figure 3.5* Behavioral Infractions, Sankey Diagram

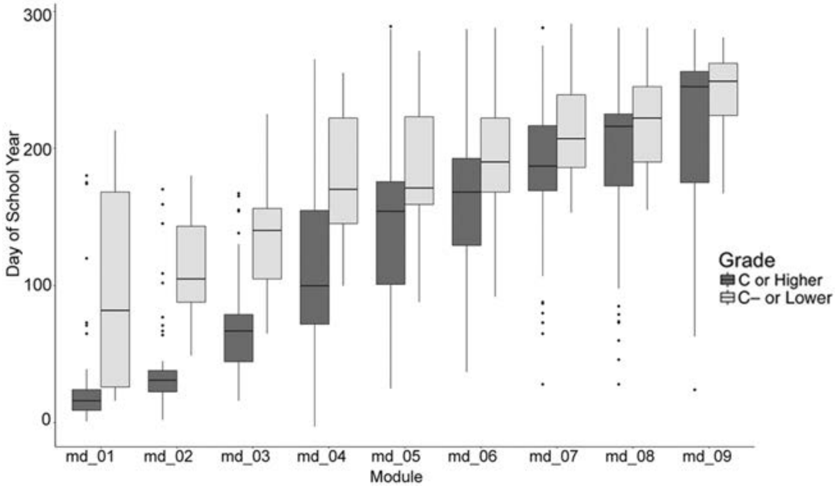


*Figure 3.6* Boxplots for Summative Assessment Passing

Figure 3.6 illustrates that there is, in fact, a dramatic difference for when students completed the first four modules based on the grade they ended up earning in the course. Interestingly, the difference between these groups of students lessens over time, which might signal that students got off track early and were not able to catch up. This particular visualization is helpful in developing a potentially useful set of features for identifying students who start slow. As evidence accumulates and our mental model for these data develops (Behrens, 1997), we can model these data to better understand the nature of the relationship between falling behind and earning a particular course grade. Along with developing models that can help us to better understand relationships among features and between features and success in the course, we can use these data to predict which students are likely to earn a C– or lower.

## Model

In Chapter 2, we introduced two general types of modeling approaches: unsupervised and supervised learning. Unsupervised learning, or structure discovery, algorithms are useful for understanding relationships among features in a dataset. Supervised learning models are different in that they can be used to quantify relationships between features and a known outcome. In this section, we expand upon these earlier distinctions by highlighting two further differences within supervised learning approaches: inference and prediction (Breiman, 2001; James, Witten, Hastie, & Tibshirani, 2013). The key difference between inference and prediction is the degree to which a researcher uses a model to interpret the relationships among features and an outcome—inference—or whether a researcher uses a model to make predictions or classifications—prediction. At the level of a project, one can use multiple structure discovery, inference, and prediction methods.

As an example of modeling focused on inference and using digital learning environment data, Krumm, Beattie, Takahashi, D'Angelo, Feng, and Cheng (2016) analyzed data from across multiple community colleges to identify patterns in students' LMS use that could drive interventions on the part of community college instructors. The purpose of the project was to develop a better understanding of students' online learning behaviors and how they related to students' performance in the course. For example, Krumm and colleagues identified the importance of students reading, practicing, and engaging in assessment activities all within the same online session as well as a negative relationship between the number of sessions students logged where they ended a session on a low score without engaging in any follow-up activity. These findings helped in building a better understanding of students' online learning behaviors. Hierarchical linear modeling (Raudenbush & Bryk, 2002) techniques were used to estimate the relationships between features that captured the behaviors previously cited and known outcomes (i.e., performance on an end-of-course exam). Particular care was given to the various assumptions that needed to be met when using linear models for statistical inference.

There are many supervised learning approaches that are not appropriate for inference as they offer limited windows into relationships among features and an outcome. Most supervised learning approaches that support inference, not surprisingly, have a history in statistics and most approaches that are geared toward prediction and classification stem from software engineering, machine learning, and data mining.

If Krumm and colleagues (2016) were less interested in asking whether or not the type of sessions students logged were related to their grades and more interested in, for example, knowing whether the number of certain session types could be used to predict students' likelihood of passing a course, then those analyses could be considered geared more toward prediction, and in particular, classification. When known outcomes are categorical, then the particular supervised learning task is referred to as classification. When the known outcome is numeric and continuous in nature, such as a test score, then this learning task is referred to as regression. The term *regression* can take on different meanings across inference and prediction uses. From a statistical, or inferential perspective, regression denotes a family of models that can be used on either categorical or continuous outcomes. Perhaps most confusing to newcomers or to researchers steeped in either inference or prediction are the ways in which specific models, such as logistic regression, can be used for either inference or classification.

Thus, a researcher's overall purpose can affect the models one is likely to use. When prediction is the focus—either regression or classification—relationships among features and the known outcome are important only insofar as they help one know what is likely to happen in the future or on *unseen* data. With this as an overarching purpose, a researcher interested in prediction may deploy multiple different algorithms in order to generate the best predictions. There are hundreds of different kinds of algorithms a researcher can use, from linear models, decision trees, support vector machines, and *k*-nearest neighbors, to name but a few examples (see Witten et al., 2017 for descriptions of these and other models). Many supervised modeling approaches fall within broader families based on their assumed relationships between an outcome and features. Two of the broadest families include *parametric* and *non-parametric* methods (James et al., 2013). Parametric methods rely on specifying a particular function at the outset of modeling, such as a linear or logistic function. Specifying a functional form at the outset has multiple benefits—most notably, it makes interpreting relationships easier. Non-parametric methods, on the other hand, make no assumptions about the functional form of relationships between features and an outcome. These more flexible models offer many benefits, but key tradeoffs include a lack of interpretability, the need for more data, and over fitting. Over fitting means a model is idiosyncratic to a specific dataset and makes predictions based on the noise in the data rather than the underlying function that generated that data.

There are a great many details that come into play when modeling data. Purpose, as we described previously, is crucial. Structure discovery helps in identifying patterns within one's data or reducing the overall dimensionality in one's data when the model is not being trained against a known outcome. Inference involves paying particular attention to the specific relationships between features and an outcome. Lastly, prediction relies on a known outcome where the ultimate task of a researcher is to find the right combination of features and an algorithm to predict or classify unseen data. In what follows, we delve more deeply into each modeling purpose. It is important to recognize that many of the details and complexities that surface with each purpose will not be addressed. The purpose of the remaining sections, as with the two data-intensive workflow steps described previously, is to help in organizing and thinking about particular steps in modeling data within the broader data-intensive workflow.

### *Structure Discovery*

There are multiple types of unsupervised learning algorithms that have been used by education researchers. Baker and Inventado (2014) outline how educational researchers have used approaches like *K*-means and hierarchical cluster analysis to *group observations* within one's datasets as well as principal components and factor analysis to *reduce the dimensionality* in one's dataset. Many of these approaches have metrics that can be used to identify which clusters or components best approximate one's data. However, without a known outcome that can be used to supervise the learning of the algorithm, judgments regarding the structure in one's data often remain subjective. Applying structure discovery algorithms can be an end in its own right; they can also be used as exploratory tools (James et al., 2013). On the one hand, unsupervised models can reveal unique patterns in one's data that can inspire new features and hypotheses, and on the other hand, principal components can directly reduce the dimensionality in a dataset that is used to build a predictive model.

Clustering algorithms are popular unsupervised learning approaches. In general, these algorithms try to organize observations, based on selected features, into groups that are similar to one another but discernibly different from other groups. Clustering approaches differ in the ways of quantifying closeness among observations and differences between groups of observations. Hierarchical cluster analyses recursively group similar observations; non-hierarchical algorithms like *K*-means clustering, on the other hand, requires a pre-specified number of groups whereby the algorithm maximizes similarity within clusters and diversity between clusters.

In education, Amershi and Conati (2009) applied unsupervised learning algorithms to data collected from students working in an environment designed to teach students how to understand artificial intelligence algorithms through animation. They used *K*-means cluster analysis to identify strategies and behaviors, such as stopping to self-explain, used by more and less successful learners. Using these techniques, they found that students who paused and engaged in self-explanation performed better on subsequent learning tasks. Other examples of using clustering with educational datasets include studying the strategies students use to solve problems in a computer-based game (e.g., Kerr & Chung, 2012), grouping similar texts written in response to questions on an assessment so that human scorers can more efficiently grade the quality of student writing (Brooks, Basu, Jacobs, & Vanderwende, 2014), and surfacing patterns that help teachers make sense of students' use of digital learning resources in order to inform subsequent course design decisions (e.g., Merceron & Yacef, 2005). Bowers (2010) used hierarchical cluster analysis algorithms to group students based on their course grades across multiple years in school to identify groups of students who eventually dropped out of high school. Lee, Recker, Bowers, and Yuan (2016) used hierarchical cluster analyses to group students based on features that characterized the way they interacted with an LMS. Using this approach, Lee and colleagues illustrated the relationship between regularly interacting with the LMS and succeeding in the courses they studied.

As we continue to work with the case high school, we are able to identify differences among groups of students based on the grade they earned and when they passed a module's summative assessment. Not well captured in Figure 3.5 are students' individual patterns for when they first passed a summative assessment; each module is treated as its own, independent grouping of days of the year. While there are multiple ways to explore students' individual patterns, we can use an unsupervised learning algorithm, agglomerative hierarchical clustering, combined with a heatmap visualization to identify groups of students based on when they completed a summative assessment. Certain packages in R, for example, make it possible to not only cluster and visualize one's data, but each observation can also be annotated. Because we are working with a dataset that has known outcomes, we can annotate each observation with whether or not a student earned a C– or lower (**Not C**). The word annotation is important because unlike a supervised learning approach, the known outcome is not training the algorithm it simply follows an observation based on the group that the observation is placed into.

Each row in Figure 3.7 represents a student and the day of the year that a student passed the summative assessment for a module. Each day of year measure can be re-scaled into standard deviation units where 0 represents the overall average for a module as to when students completed the summative assessment. Light gray cells represent students who completed an assessment earlier than their peers and black cells represent students who completed an assessment much later than their peers for

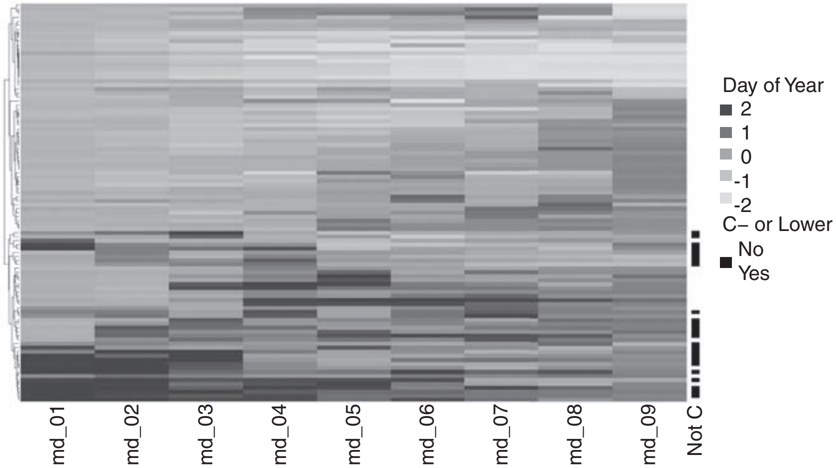


*Figure 3.7* Hierarchical Cluster Analysis With Heatmap

a given module. The clustering algorithm grouped observations based on their similarity across modules. Looking at the lower-left corner of Figure 3.7, we can see that there are students who are like one another across all modules. What is distinctive about this grouping is how similar they are on when they first passed the assessment for the first four modules. The annotation toward the right-hand side of the visualization identifies students who earned a C– or lower with a black mark in the **Not C** annotation column. The cluster of students toward the bottom-left of the figure are also those who tended to earn a C– or lower. The small cluster of students toward the middle of the figure who are also marked as **Not C** had later passing dates but much less consistently beyond module 1 (**mod_01**) as compared with the lower-left cluster, among other differences across all modules.

### *Inference and Prediction*

As described previously, inference and prediction are different purposes for using supervised learning techniques. A study by Marbouti, Diefes-Dux, and Madhavan (2016) is illustrative of using supervised learning approaches for prediction. These researchers sought to develop a predictive model that could be used to identify students at risk of failing a first-year, college engineering course using only performance data that were available to course instructors as they taught the course. For their project, Marbouti et al. (2016) were less concerned about interpreting the individual relationships between students' individual academic performances and their overall success in the course, i.e., a question of inference. Instead, these researchers were more concerned with identifying the best combination of features and one or more algorithms that could be used to predict which students are not likely to succeed in the course as early as possible in the course. In the end, Marbouti et al. used an ensemble approach, which involves training a sequence of algorithms to identify whether a student is not likely to succeed in the course, and using all of the multiple classifications made to render a final classification.

Building a predictive model is an iterative process that is somewhat distinct from building an inferential model. In broad terms, when building and evaluating an inferential model, a researcher uses all of the available data, is concerned with the degree to which certain assumptions associated with a modeling approach are met, and evaluates the model by examining residuals, which are observations' deviations from the predicted relationship between a feature and an outcome. Predictive modeling can include some of these elements depending upon the specific model used (i.e., a logistic regression model); however, these and additional steps undertaken for predictive modeling purposes are framed as *training* and *testing*.

Training involves applying a model to a subset of one's data, selecting appropriate features, and engaging in model-specific parameter tuning, which involves refining user-specified values, such as the maximum number of nodes allowed along the longest path for tree-based model describe later (Kuhn & Johnson, 2013). The basic idea behind training and testing is to avoid *over fitting* and to test the generalizability of a model, i.e., its performance on data that were not used to train the algorithm in the first place. One approach is to train a model on a feature set from, for example, 80 percent of one's original dataset. Using this subset of one's original data, training a model also involves selecting the best set of features for predicting or classifying the known outcome. To help in determining which features to include, some algorithms, such as decision trees, select the best features based on the way they partition data and the parameters set by a researcher. Training a model can further involve algorithmically recreating the test–train split described earlier but only on the training dataset—referred to as cross-validation. Cross-validation involves breaking a dataset into *k* number of sub-samples, holding out one sub-sample, and using the remaining data to train a model that is then tested on the held-out sample. This process happens for each held-out sample and can be repeated a desired number of times. After selecting features, selecting an algorithm, and tuning parameters, the best performing model is then applied to the 20 percent held-out data, which is referred to as the validation dataset. In some cases, one can use an out-of-sample validation set, such as another academic year's data, to evaluate a model trained on data from one academic year. For example, Knowles (2015) trained a high school dropout classifier using one academic year's data and then tested the trained model and feature set on the following year's data. This approach can work well if there is enough similarity in terms of available data and the underlying activity across the two academic years.

Prediction methods, with variations on the themes described previously, have been used to predict and classify a variety of different educational outcomes, including not only course performance but also higher-inference constructs such as inquiry science skills (Sao Pedro et al., 2011) and affective states (D'Mello et al., 2008). Gobert, Baker, and Wixon (2012) sought to predict when students were disengaged while using an ITS. To do so, trained researchers coded clips of students' ITS use and noted each time a student appeared to be off task. The clips coded by trained researchers were then used as known outcomes on which models were trained. Thus, in recent years, it is possible to automatically detect a range of student engagement and meta-cognitive variables that can be used to better respond to the needs of an individual learner, which is often referred to as personalizing learning (Pea, 2014). In fact, many questions explored in the educational data mining space center on predicting small-scale behaviors of learners to drive more personalized, technology-driven adaptations.

Following the unsupervised learning analyses in our case high school, we have added yet more evidence to the idea that students starting the course slow may be a useful organizing idea for creating features that can be used to understand and predict students earning a C– or lower in the course. For the purpose of inference, we may wish to know the degree to which falling behind affects the likelihood of a student earning a C– or lower. For the purpose of prediction, we may supply an algorithm multiple features, such as the same features used in the hierarchical cluster analyses, i.e., day of year when student first earned greater than 60 percent (see Figure 3.7), and let the algorithm determine the model that best identifies students likely to earn a C– or lower. For this task, we can use a conditional inference tree model. Decision trees classify observations by partitioning and sorting them from the *root* of a tree out to the *leaves*, which represent the values of a known outcome, such as "C– or lower" or "C or higher." The root represents the best attribute for classifying observations, and an observation is classified by starting at the root of the tree and following it out to a corresponding value for a known outcome. Figure 3.8 illustrates our predictive model for students earning a C– or lower. The best predictor was the day of year that a student completed the summative assessment for the second module (**md_02**). The value for this attribute, and in the case of this model, the root, was the 45th day of the year.

After training the model, we can then apply it to our testing dataset. There are multiple metrics that can be used, especially when using classifiers, for assessing the overall performance of a model. Many of these metrics are based on the idea of a confusion matrix, or contingency table. For dichotomous outcomes, the ground truth is either "true" or

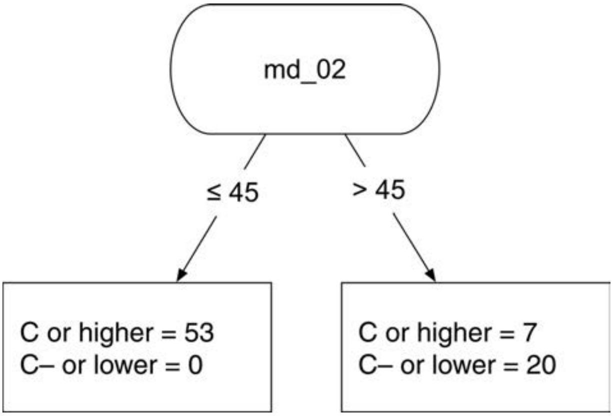


*Figure 3.8* Conditional Inference Tree Plot From Training Data

"false," meaning that what is being classified is coded as "true" because it represents known outcomes. In the case of identifying the likelihood of students earning a C– or lower, a positive coding is "yes." Combined with a model's predictions, we can identify a variety of configurations and proportions across, for example, true-positives and true-negatives (see Table 3.1). In identifying true-positives, we want to reduce the number of false-positives, students who have a value for a certain attribute and are thus predicted to C– or lower but in reality earned a C or higher (see Bowers, Sprott, & Taff, 2013). In building this predictive model, we created an 80/20 split in our original data and trained the conditional inference tree model on 80 percent of the data using ten-fold cross-validation, which partitioned the data into 10 equal sub-samples whereby one sample was held out and tested against the remaining data, one time for each sub-sample. The best model from this process (see Figure 3.8) was retained and then applied to the 20 percent of the original data that were held out as validation data. Table 3.2 provides the results for our validation data. Of the seven positive cases, i.e., those students who earned a C– or lower, our model identified five of these cases, and two cases based on the value of the greater than the 45th day were false-positives, meaning that they first earned a passing grade later in the year but earned a C or higher. All in all, this model may prove useful for next year's teachers as a way of identifying students in need of support.

*Table 3.1* Generic Confusion Matrix

| | | *Ground Truth* | |
|---|---|---|---|
| | | *Yes* | *No* |
| **Prediction** | **Yes** | True-positive<br>(TP) | False-positive<br>(FP) |
| | **No** | False-negative<br>(FN) | True-negative<br>(TN) |

*Table 3.2* Case School Confusion Matrix

| | | *Ground Truth* | |
|---|---|---|---|
| | | *C– or lower* | *C or higher* |
| **Prediction** | **C– or lower** | 5<br>(TP) | 2<br>(FP) |
| | **C or higher** | 0<br>(FN) | 13<br>(TN) |

## Conclusion

Any text that purports to cover methods, especially at a high level like we have done in this chapter, is bound to leave many topics out. Thus, there are obviously many methods that we did not address. For example, social network analysis offers a powerful set of tools for understanding relationships and communication patterns among, for example, individuals in an online learning environment (e.g., Anaya, Boticario, Letón, & Hernández-del-Olmo, 2015). Text mining and natural language processing models are nearing ubiquity and being used, for example, to understand MOOC completion (Crossley, Paquette, Dascalu, McNamara, & Baker, 2016). And for the educational data mining research space, in particular, we did not address the vast array of approaches related to student knowledge modeling, such as Bayesian Knowledge Tracing (Corbett & Anderson, 1995). Moreover, there are many statistical, i.e., inferential, models that we did not discuss, such as longitudinal growth models, mixture models, hazard models, or mixed-effects models (e.g., hierarchical linear models)—as well as psychometric approaches like Item Response Theory. While there are many statistical approaches that we did not discuss, there are just as many machine learning techniques that were not introduced. As of this writing, the caret package, which is a package used in R that contains many machine learning tools, had over 232 available models.

In spite of these and many more omissions, our goal in this chapter was to highlight the intuition and thinking that goes into wrangling, exploring, and modeling data. For researchers, there are many methods and models to choose from in analyzing data from digital learning environments, administrative data systems, and sensors and recording devices. These data, along with specific models and analytical techniques, are inputs into a data-intensive research workflow that culminates in data products that are communicated to an interested audience. This chapter paid particular attention to the wrangling, exploring, and modeling steps in the workflow and used a hypothetical case to highlight how an entire project can play out across these steps. Before using the types of methods described previously, it is important that the right people have been brought together to ensure that the best questions are being asked that will benefit both researchers and practitioners. This upfront work can not only lead to better analyses, it can make the work that follows easier too. For example, when working in a partnership, practitioners are more likely to take up a potential change idea derived from an analysis if they had a hand in shaping the original analysis. Chapters 6 and 7 take on these issues by explicitly focusing on what needs to be in place as well as how to engage in partnership-driven data-intensive research. In the next two chapters, we focus on issues around data privacy that make data-intensive research possible in the first place. Following that, we describe the history of different educational research approaches that have played a role in shaping the current data-intensive research landscape in education.

## References

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help helps, but only so much: Research on help seeking with Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education*, *26*(1), 205–223.

Amershi, S., & Conati, C. (2009). Combining unsupervised and supervised classification to build user models for exploratory learning environments. *Journal of Educational Data Mining*, *1*(1), 18–71.

Anaya, A. R., Boticario, J., Letón, E., & Hernández-del-Olmo, F. (2015). An approach of collaboration analytics in MOOCs using social network analysis and influence diagram. *The 8th International Conference on Educational Data Mining (EDM 2015)* (pp. 492–495). Madrid, Spain.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of Intelligent Tutoring Systems. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531–540. Macei´o, Alagoas, Brazil.

Baker, R. S. J. D., Inventado, P. S. (2014) Educational Data Mining and Learning Analytics. In J.A. Larusson, B. White (Eds.) *Learning Analytics: From Research to Practice.* Berlin, Germany: Springer.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131–160.

Bowers, A. J. (2010). Analyzing the longitudinal K–12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment Research and Evaluation*, *15*, 1–18.

Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity and specificity. *The High School Journal*, *96*(2), 77–100.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Brooks, M., Basu, S., Jacobs, C., & Vanderwende, L. (2014). Divide and correct: Using clusters to grade short answers at scale. In A. Fox, M. Hearst, & M. Chi (Eds.), *Proceedings of first (2014) ACM conference on learning @ scale (L@S '14)* (pp. 89–98). New York: ACM.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278.

Crossley, S., Paquette, L., Dascalu, M., McNamara, D., & Baker, R. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the 6th International Conference on Learning Analytics and Knowledge*, 6–14. Edinburgh, UK.

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology & Society*, *17*(1), 17–28.

D'Mello, S. K., Craig, S.D., Witherspoon, A. W., McDaniel, B. T., & Graesser, A. C. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction, 18*(1–2), 45–80.

Feng, M., Heffernan, N. T., & Koedinger, K. R. (2009). Addressing the assessment challenge in an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI Journal)*, *19*(3), 243–266.

Gobert, J. D., Baker, R. S. J. D., & Wixon, M. B. (2013). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist*, *50*(1), 43–57.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (1st ed.). New York, NY: Springer.

Kerr, D., & Chung, G. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining*, *4*(1), 144–182.

Knowles, J. E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining*, *7*(3), 18–67.

Kotu, V., & Deshpande, B. (2014). *Predictive analysis and data mining: Concepts and practice with RapidMiner*. Waltham, MA: Morgan Kaufmann.

Krumm, A. E., Beattie, R., Takahashi, S., D'Angelo, C., Feng, M., & Cheng, B. (2016). Practical measurement and productive persistence: Strategies for using digital learning system data to drive improvement. *Journal of Learning Analytics*, *3*(2), 116–138.

Krumm, A. E., Boyce, J., Gassman-Pines, A., Bellows, L., & Podkul, T. (2017). Inequality and educational datasets: New opportunities to explore critical issues through cross-agency collaborations. Poster presented to the *Annual Meeting of the Association for Education Finance and Policy*. Washington, DC.

Kuhn, M. & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

Lee, J. E., Recker, M., Bowers, A. J., & Yuan, M. (2016). Hierarchical cluster analysis heatmaps and pattern analysis: An approach for visualizing learning management system interaction data. In T. Barnes, M. Chi, & M. Fen (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)*. Raleigh, NC.

Loeb, S., Dynarski, S., McFarland, M., & Reardon, S. (2017). *Descriptive analysis in education: A guide for researchers*. (NCEE 2017–4023). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20174023/

Madhavan, K., & Richey, M. C. (2016). Problems in big data analytics in learning. *Journal of Engineering Education*, *105*(1), 1–9.

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, *103*, 1–15.

Merceron, A. & Yacef, K. (2005). Educational data mining: A case study. In C.-K. Looi, G. McCalla, B. Bredeweg, and J. Breuker (Eds.), *Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology* (467–474). IOS Press, Amsterdam, The Netherlands.

Murphy, R., Snow, E., Mislevy, J., Gallagher, L., Krumm, A. E., & Wei, X. (2014). *Blended learning report*. Menlo Park, CA: SRI Education.

Paquette, L., de Carvalho, A. M. J. A., Baker, R. S., & Ocumpaugh, J. (2014). Reengineering the feature distillation process: A case study in the detection of gaming the system. In *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 284–287). London.

Paquette, L., Rowe, J., Baker, R. S., Mott, B., Lester, J., DeFalco, J., Brawner, K., Sottilare, R., & Georgoulas, V. (2015). Sensor-free or sensor-full: A comparison of data modalities in multi-channel affect detection. In *Proceedings of the 8th International Conference on Educational Data Mining* (93–100). Madrid, Spain.

Pea, R. (2014). *The learning analytics workgroup: A report on building the field of learning analytics for personalized learning at scale*. Stanford: CA: Stanford University Press.

Peng, R. D. (2016). *Exploratory data analysis with R*. Leanpub. Retrieved from https://leanpub.com/exdata

Penuel, W. R. & Gallagher, D. (2017). *Creating research-practice partnerships in education*. Cambridge, MA: Harvard Education Press.

Piety, P. J., Hickey, D. T., & Bishop, M. J. (2014). Educational data sciences: Framing emergent practices for analytics of learning, organizations, and systems. In S. Teasley, and Z. Pardo (Eds.), *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 193–202). New York: ACM.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage Publications.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–291.

Sao Pedro, M. A., Baker, R. S. J. D., Gobert, J. D., Montalvo, O., & Nakama, A. (2011). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction*, *23*(1), 1–39. Montreal, Canada.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

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

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

VanderPlas, J. (2017). *Python for data science handbook*. Sebastopol, CA: O'Reilly Media.

Ventura, M., Shute, V. J., & Small, M. (2014). Assessing persistence in educational games. In R. Sottilare, A. Graesser, X. Hu, & B. Goldberg (Eds.), *Design recommendations for adaptive intelligent tutoring systems: Learner modeling* (Vol. 2, pp. 93–101). Orlando, FL: U.S. Army Research Laboratory.

Wickham, H., & Grolemund, G. (2017). *R for data science*. Sebastopol, CA: O'Reilly Media.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data mining: Practical machine learning tools and techniques* (4th ed.). Cambridge, MA: Morgan Kaufman.

Chapter 4

# Legal and Ethical Issues in Using Educational Data

Sensitivity over use of personal data in education can be understood within the broader context of public sensitivity over the collection and use of data more generally. Some might argue that big data are nothing new, because scientists and demographers have been collecting large amounts of data about people for some time. But today's data collections capture much more information about people's day-to-day and moment-to-moment actions, oftentimes with a corresponding feeling of invasion of privacy. The machine learning expert Hannah Wallach makes a distinction between big datasets arising in the sciences (such as physics and astronomy) and those involving individuals' interactions with systems that are now causing such concern (Wallach, 2014). She argues that because the latter type of big data involve statistical information on social, economic, and human behavioral trends—in other words, that it's about *people* and how they act, what they like, and who they are—these *social* datasets are what give us pause.

In day-to-day life, anyone who has engaged in purchasing over the Internet has experienced the annoyance of having data collected to improve marketing that results in targeted ads that pop up during any Internet browsing for months. Many of us, though, are willing to give up some privacy—such as anonymized location information—for a benefit—less sitting in traffic. A 2015 Pew Research survey found that Americans have different and evolving perceptions of the risk vs. benefits of releasing personal information based on the location where the data are collected. For example, 54 percent of those surveyed rated it acceptable, in light of recent thefts of personal belongings, for a workplace to install high-resolution security cameras that could perform facial recognition, even if the company subsequently chose to use video footage for rating employee attendance. In contrast, only 27 percent of respondents rated as acceptable home temperature control sensors that could save energy by adaptively adjusting temperatures based on movements around the home if they result in the company having information about when people are in various rooms in the home and when they move from room to room

(Rainie, 2016). Still, Pew found people willing to give up data for free services, such as email in exchange for targeted advertisements, and resigned this tradeoff as a part of modern life, while also desiring better legal remedies, such as better laws requiring disclosure of how information is collected and used (Rainie & Duggan, 2016).

One way in which organizations that collect large amounts of personal data are responding to these concerns is by enabling personal data downloads. You can download your Google browsing history via Google Takeout, for example, and your messages, connections, and contacts from LinkedIn. From Facebook, you can download your profile—birthday, gender, current city, hometown, family, education, employers, favorite restaurants—a timeline of all posting activity, ads you have clicked on, private messages sent, friends deleted, religious views, the IP addresses from which you have logged into or out of Facebook, and facial recognition data. Data downloads can show consumers the scope of data being collected and maintained by companies but don't reveal how those data are used and shared across company boundaries. This lack of transparency has led to calls for more control over not just the collection of personal data but also its *use* (Harvard Business Review Staff, 2014).

Approaches to addressing privacy concerns range from legal protection to cultural and social norms governing data collection and use. In the U.S., federal and state laws have been enacted that govern how research, educational, and commercial enterprises should handle personal data, and these laws are updated frequently to reflect new concerns about big data. As we describe in the remainder of this chapter, privacy concerns are driving legislation that could have a dampening effect on educational research. Sometimes there are competing and seemingly incompatible values, as when the ideal of open science competes with the individual's right to have past behaviors be forgotten. Other times, there are misinterpretations, misperceptions, or real or imagined harm. In this section, we provide background on privacy and related topics, legality, and ethics in using education data.

According to a report from the National Science and Technology Council (NSTC) in the Office of the President under the Obama administration (NSTC, 2016), context is imperative in understanding privacy. Certain circumstances warrant the sharing of private information. Private information can be and is regularly accessed by *authorized* users (e.g., bank account information by a credit reporting agency) under appropriate circumstances. Information security and privacy controls make sure that only authorized users access private data. In the next paragraphs, we draw from the 2016 NSTC report to discuss concepts that will be used in the rest of the chapter: privacy, information security, and trust.

The NSTC report treats *privacy* as having to do with expectations of control of personal information relative to a context. Individuals do not expect that personal information or data will always be kept secret to themselves or never shared with others, but they share their data with organizations or individuals with the expectation that the data will be confined to that context. Health care, law, education, and religion are all areas where individuals may share information that they expect will be kept within that context and only used for specific purposes. Violations of privacy relate to uses outside of the context within which the data were collected that violate the expectations of the individual sharing the data.

*Information security* refers to standards for data storage of personal information that are implemented to prevent unauthorized access to data. Early in the adoption of online systems to manage student information, information about students was not all sent over secure channels and thus was subject to potential "eavesdropping." This is a clear example of a lack of information security and, in 2014, led to a proposed change in the Family Educational Rights and Privacy Act (FERPA) to add language about private companies using proper security safeguards.

*Trust* is a belief, based on a relationship, that the individual or organization to which you supply data will not willingly or unwillingly make those data available to additional parties without your consent. Trust in privacy in a specific context, then, means that individuals believe that they have control over, or knowledge of, use and disclosure of their personal information, and that they believe there are security controls in place to ensure that information cannot be accessed by unauthorized users. Disclosure occurs when what was shared within one context—with presumed controls—is shared in other contexts, and the information owner thus loses control. In the case of big data, privacy concerns about personal information are heightened because, as we have described, they are perceived to be more personal and somewhat intrusive. In data collected for educational use and research, questions arise about the reasonable privacy expectations of students and parents. What and whom do they trust? What are the potential risks of disclosure?

Regulating the control of information requires a precise definition of the term "information." For educational purposes, the U.S. Department of Education has created a Privacy Technical Assistance Center (ptac.ed.gov) to help state and local agencies interpret privacy laws that apply to education and to online protections for children. According to the PTAC website, personally identifiable information (PII) in education records encompasses both direct identifiers (e.g., student's name or identification number, address, Social Security number, telephone number, email address) and indirect identifiers, such as date or place of birth, race, religion, and activities. Sometimes, connecting data together makes it PII, as when one can distinguish or trace an individual's identity because he is the only student at the school from a particular zip code or of a particular ethnic background. In simple terms, personally identifiable information consists of data that, directly or indirectly, can be used to identify a particular person.

Privacy concerns from parents and privacy advocates have risen as companies have begun supporting educational entities through data warehousing (see Box 4.1) or, more recently, by offering online educational experiences that may collect PII or other sensitive information. Such concerns have been heightened by revelations that government agencies have been spying on U.S. citizens and by the data breaches at large retailers (Bulger, McCormick, & Pitcan, 2017). Detailed learning process data collected in the form of system log data and performance on assessments could be considered sensitive information if that information has the potential to cause embarrassment, unfair treatment, or other harm if released. Derived or inferred data, such as classifying a student as a potential dropout on the basis of predictive analytics, are also likely to be considered sensitive. Parents may be comfortable with an education system collecting such data to improve student learning but may become concerned when data leave the school or district's control, especially if they go to private companies.

Generally speaking, the potential harm that may come from release of such information arises when it is associated with PII. Accidental or malicious release of data can come from data leaks (e.g., school officials accidentally disclosing student data), hacks, or misplaced or forgotten data. There may be failure stemming from the absence of a plan for data management when companies supporting schools, school districts, or state education agencies go out of business (Molnar, 2014). Potential harms that could follow such releases include, in the education sector, emotional responses (e.g., embarrassment), bullying, and damage to reputation. However, while reports of breaches and misuse of data can be found (e.g., the Privacy Rights Clearinghouse), reports of actual harms from release of education data are more difficult to find. Many of the incidents listed at the Privacy Right Clearinghouse concern release of personnel data rather than students' education data. Breaches of education data have been reported (e.g., a web-accessible plain-text database of student information from a network of charter schools in California, including name, gender, grade, and disability status; DataBreaches.net, 2015), as was an accidental release of similar information by District of Columbia Public Schools (Stein, 2016), but actual substantiated reports of harms have not emerged (Herold, January 2014). This lack of actual reported harms, however, does not mean that data security concerns are invalid; they must be taken seriously, and in many cases, there are not only ethical obligations but also legal ones, as we describe next.

## Regulations and Laws on the Use of Data in Education Research

Across the board, for any scientific research involving human subjects in the U.S., the Common Rule (U.S. Department of Health and Human Services, 2009) sets forth guidelines for ethical behavior for those who conduct such research. The Common Rule provides guidance for how research is reviewed, how one informs subjects of their participation in research, and how one obtains consent from subjects to participate in research. Consenting to participate relies upon trust that the data collector will only use the collected data for the stated purpose and carries an assumption of privacy—that any data that could be personally identifiable will be protected from disclosure.

The Common Rule, and regulations that implement it—from agencies that sponsor or conduct research with human subjects—guides the work of institutional review boards (IRBs) for the protection of human subjects at universities and nonprofit research organizations that receive federal research funding. IRBs are typically committees or individuals who receive applications for conducting research, review the applications, and make judgments and recommend alterations to research procedures based on the statute governing research with human subjects (i.e., the Common Rule) and other applicable regulations (e.g., FERPA in the case of research involving students). While the complete set of human subjects regulations and guidelines is too complex to discuss here, it is worth noting that the collection and analysis of education data can be categorized as exempt from participant consent practices if it is deemed part of normal education practice or if the data are de-identified. Such exemptions do not mean that the data are any easier to get, however; regardless of consent requirements, agreements to govern data use and security are needed between all parties sharing data.

### *Federal Legislation Specific to Education Records*

In the U.S., the legal responsibilities of educational institutions that gather, store, and use student data and the organizations they partner with to analyze those data—including universities, nonprofits, and companies—are set forth by FERPA, first enacted in 1974. FERPA imposes these responsibilities as a means of protecting parents' and students' rights. It gives parents and eligible students the right to review students' educational records and requires schools to obtain their consent before disclosing any student information contained in those records to third parties—except in case of a specified set of exemptions. In other countries, e.g., in the European Union, educational data are governed by broader and more general privacy legislation that covers any entity that controls personally identifiable data and requires researchers to "obtain personal unambiguous consent before data can be processed" (Har Carmel, 2016, p. 8). In the U.S., all schools receiving federal funds must comply with FERPA, and even if no federal funding is received, it behooves any organization to comply. Increasingly, as we will discuss later, states are taking on privacy legislation also, sometimes passing laws that are more stringent than the federal regulations.

The stated default for FERPA is to obtain consent for data collection and use, but the regulations include exceptions to allow educational institutions more efficient operations. One exemption relates to data collected on the operation of the education enterprise, including assessments, attendance data, library records, and a host of other records. Educational entities—schools, school districts, and colleges—are expected to gather and analyze administrative data for the purpose of improving their services to students. Even though these administrative data contain PII, such routine data collection does not require notification or parental or student consent under FERPA.

FERPA also applies in higher education, and the same educational purposes exemption applies. But suppose rather than just improving the educational enterprise, someone wants to collect or explore the data in a systematic way to further knowledge. In this case, the data become research data and, if not de-identified, fail to qualify for exemption under FERPA. To take an example, suppose faculty at postsecondary institutions have been using data collected during the courses they teach to improve their teaching. They may be working with an office at their university that supports online learning or helps faculty teach more effectively. Such data collections generally are not subject to FERPA requirements. However, if faculty and staff collecting the data want to share their findings and participate in scholarship on teaching and learning, they need to view their work as research and conform to the regulations for disclosing use of education data as well as those protecting human subjects. For example, if they wish to address a research question such as whether their teaching approach differentially affects certain student subgroups, such as English language learners, and make their results public, they enter the realm of research and have to receive approval from an IRB.

FERPA supports parents in protecting their child's PII. It also supports educators and administrators by outlining what data they are permitted to collect and use without notifying parents, and how they can extend that permission to disclose data to authorized users, which may include researchers. FERPA allows its "School Official and Audit/Evaluation" exemption for use of PII to be extended to cover third parties, including IT organizations. The rationale is that these organizations are (1) using the disclosed data to provide the same type of services that a school/district employee might, (2) providing a service that is controlled by the school/district (e.g., storing records in a data warehouse under contract to the district), and/or (3) using the data for "legitimate educational interests," much as a school official would. For example, FERPA allows schools and districts to use cloud-based or other external hosting services provided by private companies to host data that includes PII without notifying parents of their choice (Privacy Technical Assistance Center, 2012a). The company is also bound by FERPA, however, and will need to certify in its terms of service or other contract with the school or district that PII will not be used for unauthorized purposes. Some people have objected to outsourcing of educational functions involving data, but others have argued that private companies have better information security practices in place than districts ever could, and that this practice can prevent accidental disclosure or disclosure by malicious actors.

FERPA limits what an authorized agent can do with PII student data. For example, there have been concerns about private companies using student data for advertising purposes. We are all familiar with data collection during online browsing being used to target advertisements. Since advertising is not an educational activity, it is forbidden by FERPA (Data Quality Campaign, 2015). Understanding what this restriction means in practice has become clearer over time. For example, Google Apps offers mail and document management in the cloud, and because students and teachers often adopt these apps for their own personal use, it's easy enough to see how the education version, Google Apps for Education, could migrate to the classroom without much oversight by a school or district administration. But consider the year 2011, when Google Apps for Education was using the same mail scanning system as Google Apps did for non-education mail management services. Google offered the same privacy statement for Google Apps for Education users as it did for consumers. As a result, Google was scanning student emails, as it did consumer emails although it was not displaying advertisements to users of its Apps for Education. This scanning of student data was unexpectedly uncovered during a lawsuit about Google's scanning of consumer use data (Herold, 2014). Now Google no longer scans emails in Google Apps for Education.

Educational research is generally covered under the studies section in the FERPA statute (Privacy Technical Assistance Center, 2012b) as well as being regulated by IRBs. When researchers use educational data, they need to comply with FERPA rules around handling PII, including using it for research purposes, keeping it secure, and not disclosing it outside of the set of authorized users. Written agreements between researchers and schools/districts, called data-use agreements (DUAs), can outline exactly which administrative PII data are to be given to the research organization, how the data will be used, and when the data will be destroyed. Non-administrative data with PII collected by the research team, covered under IRB, may require parental consent and student assent.

While the recommended type of data to share within research–practice collaborations is de-identified data, which fall outside of FERPA (Privacy Technical Assistance Center, 2014), this practice can be unwieldy because it shifts the burden of aligning data collected from different sources (i.e., matching records pertaining to the same student that come from different sources) from the research organization to the school or district. It is also possible that the merging of records results in having enough information about individuals to re-identify them (Leichty & Leong, 2015) or that the loss of the identifiers weakens the analysis (Daries et al., 2014). Obtaining consent from parents to use PII to make the job simpler is not a useful fallback because it brings up issues when some parents opt out, potentially damaging the representativeness of the sample. As online educational companies fill the K–12 market and researchers combine traditional educational research data collection and data from online learning systems, new laws may push for parental consent.

In summary, PII data may be collected, stored, and used by a third party for legitimate educational purposes under the FERPA "school official" exemption or through written consent. De-identified data can be used for a broader set of purposes, including research and product improvement. Research that utilizes PII is permitted under FERPA as long as the PII is not disclosed to parties outside of the researchers who have an interest in the data, the information is protected, and the data are destroyed when no longer needed. We note that this last requirement is incompatible with the move toward open science, and expect further evolution of guidance around data sharing and destruction in the years to come.

Finally, we note that these federal laws constitute the floor for protection of student data, and learning analytics researchers should be aware that revisions to FERPA have been suggested and likely new amendments will be introduced. "Ceilings," as we shall see next, are increasingly coming from states that are making stricter data privacy laws.

### *State and Local Legislation*

Recently, states have been stepping in to address what they see as holes in federal laws protecting student data. State laws around use of education data are expanding—and sometimes confusing—and they may create unintended consequences for educational research and improvement. While the pace of state student data privacy legislation has slowed down from its highs in 2014–15, privacy remains an active issue in state legislatures. In 2016, 34 states introduced 112 bills addressing student data privacy and 15 states passed 18 new laws (Data Quality Campaign, 2016). Many of these laws address transparency and contain provisions specifically giving parents greater insight into the data that are collected on their children. Districts and higher education institutions may interpret such provisions to mean that applications to conduct research studies in their schools must contain more specific information about the data to be collected and how those data will be used, stored, accessed by parents upon request, and potentially deleted upon request. They may also demand written consent from parents or from students over age 18.

State data privacy laws and local policies are changing, so readers are advised to keep track of changes at their local school board level and through state-focused news from organizations such as the National Association of State Boards of Education's project on education data privacy, the Data Quality Campaign, the Future of Privacy Forum, and Common Sense Media.

In our work, when we collect student academic records from a school or district, we execute a data use agreement (DUA) that specifies how we will use, protect, and dispose of data provided for analysis. Researchers need to be clear what counts as student records or administrative data. Suppose researchers have created an assessment of student learning that is administered as part of the research. If a teacher uses student performance on that assessment as part of a student's grade, it becomes part of the student's academic record and therefore becomes administrative data and may need to be covered by a DUA. IRBs can help sort through such situations. Within the DUA, we also stipulate that we will not disclose the data to any third party or subcontractor without notification and approval. When we work with an online learning services provider to obtain platform-generated log data, we ask the provider to execute a DUA with the school to notify it that the provider is supplying student-level data to us. Data use agreements make clear how the data will be used, and it behooves researchers to explain this clearly. Researchers need to be aware of whether the DUA signed with the local education agency has restricted data use in some ways.

Careful descriptions of what data will be collected or shared and how those data will be used need to be included in DUAs. Otherwise, use of de-identified data for product improvement might not be covered by the agreement (U.S. Department of Education, 2014). Another consideration that should be covered in DUAs and parental consent forms is that data collected for a particular research study might be required to be archived for future use by other researchers under the research funder's open science requirements.

More districts are establishing formal application processes to obtain permission to conduct research in their schools. Some are requiring formal consent for any staff or teacher survey as well as for student data collections, irrespective of any ruling on written consent requirements by an IRB. In a future of stricter privacy rules, some districts or states may require written parental consent even for uses of administrative PII data. Additional work is required, usually on the part of school staff, to distribute consent forms to parents and track which students have returned forms with parent signatures. The typical method for distributing such forms by sending them home with students is vulnerable to loss both on the way to parents and on the way back. Further, school staff need to deal with the issue of what students without consent will be doing while students with consent are engaged in research activities. Parental consent requirements thus pose a challenge for educational research, but this has not stopped states such as Arizona from passing legislation requiring parental consent for the collection of any personal information from students.

Some school districts are getting out ahead of privacy issues and adopting their own privacy, security, and Internet safety policies. For example, the Houston Independent School District (ISD), the largest school district in Texas with over 200,000 students, has created a method for district staff to rate the data practices of web applications when they are selecting digital learning tools and applications. Working with the Council of Great City Schools, Common Sense Media, and the Future of Privacy Forum, the Houston ISD developed a rubric to evaluate whether the digital learning resource under consideration complies with federal laws around protecting students, including FERPA, as well as whether data are to be transmitted over encrypted links, whether PII is even collected, what the application provider's published privacy policy is, and how it is disclosed, whether user data can be deleted when accounts are cancelled, and whether the site advertises and, if so, whether the ads are appropriate. Such rubric ratings, if posted publicly on the district website, can exert pressure on companies and other providers to bring their security and privacy procedures up to date.

The flowchart in Figure 4.1 captures many of the issues one needs to think about in protecting data privacy. Of course, as the figure shows, steps in this process do not obviate the need for working with an organization's IRB, working with a district or state education agency, or executing written agreements around specific data collection and analysis efforts.

## Ethical and Responsible Use of Data in Education Research

We have been talking about laws, statutes, and regulations, and as Buchanan (2015) points out, the privacy focus in education has been largely about compliance, restrictions, and technical approaches. Ethics, in contrast, concern doing what is right, whether or not it is enforced by regulations and laws. Ethics are about right and wrong, and the moral standards that distinguish between them. In this section, we outline potential ethical concerns that arise from using data in educational contexts and to make instructional decisions.

Ethical decisions involving use of big data can be characterized in terms of ethical principles for use of the data and its results—doing no harm through the use of big data; doing good; making sure that the applications and results are just, fair, and equitable; and ensuring that individuals have agency in terms of decision making or, at least, transparency into what decisions are being made about them.

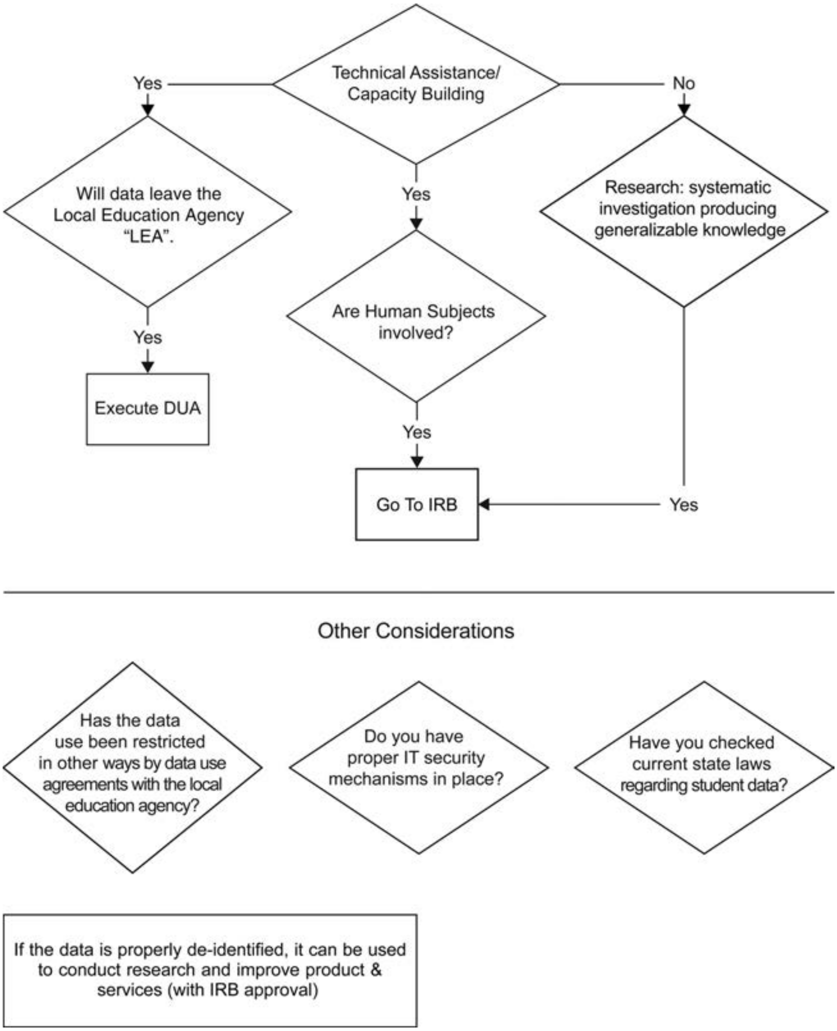


*Figure 4.1* Protecting Student Data Privacy

Examples of poor uses of data in educational contexts include targeting ads to students, not making clear what happens to data when a company goes bankrupt or is sold, making decisions about students based only on big data, and not asking for accountability in the algorithms used. More responsible uses of data include improving a product or a practice based and using multiple sources of data rather than a single metric when making consequential decisions.

Commercial organizations that collect and use data can do so responsibly by adopting the Fair Information Practice Principles from the Federal Trade Commission. These state that an organization should not collect more data than needed, should not collect or store inaccurate data, should specify how data are to be used and get permission for its use, and should get permission, when possible, before using data in new ways.

While no privacy agreement is perfect, the examples from the U.S. Department of Education for educators wishing to develop terms of service (Privacy Technical Assistance Center, 2014) and the Software and Information Industry Association (SIIA) are helpful resources that acknowledge that security and privacy practices need to explicitly address parents' and students' potential concerns. Critically, these resources make it clear that what the law requires is a *minimum* standard for what can and cannot be done with student data. SIIA also collaborated with the Future of Privacy Forum to create the Student Privacy Pledge (https://studentprivacypledge.org/). Companies signing the pledge promise to make clear how they meet federal law and regulatory guidance covering the collection, maintenance, and use of students' personal information.

Researchers shouldn't discount or dismiss the voices of stakeholders like parents, even when those voices seem to come from an antagonistic vocal minority. Methods of communication include privacy policies, consent and assent forms, and data use agreements. Schools, unfortunately, may not make, or may not feel they can make, the fine distinction between data used for research purposes and data collected by online providers for improving products.

## Transparency, Accountability, and Fairness in Algorithms

Imagine a data scientist studying students who "game the system" when learning online by looking for ways to get through the online learning assignments without really working to grasp the learning content. The data analyst may have identified a set of online behaviors—such as going directly to online quizzes without reviewing study material—that she defines as gaming, and may have demonstrated that such behaviors correlate with low scores on the course final examination. Now suppose the researcher labels students who exhibit these behaviors online as "slackers" to distinguish them from students with other online behavior profiles. Perhaps the research team brainstorms interventions to re-engage such students in the online learning activity. But what if, through a hack or incomplete or insecure practices, the names of children diagnosed as "slackers" become public? While the intention was to help all students do better, the labeling might affect how teachers interact with the labeled students in ways that are deleterious. Moreover, it may be that there are biases built into the algorithm used to identify slackers. Given such potential for harms and loss of trust, it is incumbent upon researchers engaged in data-intensive research to anticipate and take measures to avoid potential compromises to data security. Within a research–practice partnership, data confidentiality and security efforts are part of building and maintaining trust; central to ethical professional practice; and of course, necessary to be compliant with applicable federal and state laws on data use and privacy.

### Box 4.1 Contrasting the inBloom and PAR Student Data Initiatives

Perhaps the highest-profile education example of a failure to adequately address privacy concerns was the data integration effort known as inBloom. Started as the "Shared Learning Collaborative" project funded by the Gates Foundation in 2011, inBloom's intent was to build and deliver a data warehouse or storage service that could integrate data from multiple systems with a school district. The idea behind the project, which was eventually spun off into a nonprofit organization, was to implement a specification for data interoperability.

InBloom was built as "middleware" that would support data storage, merging from disparate databases, and integration into user-facing applications and reporting systems. InBloom sought to bring together, under a common and open source data architecture, legacy systems' data as well as new data coming in from, for example, learning management systems. The idea was to free data from proprietary datastores and applications. To drive development, user demand, and more widespread adoption, inBloom signed on large and prominent districts.

To many journalists and parents, this development seemed like an example of the dangers foretold by the Electronic Privacy Information Center (EPIC) in its lawsuit against the U.S. Department of Education that alleged that FERPA's laxity would allow private companies to collect and use student data. These constituents did not see inBloom as collecting data on behalf of a school official for a legitimate educational purpose: instead, inBloom was branded as a "big data broker" during the public outcry that arose when news of district adoptions was reported.

We think it is instructive to compare the eventual shuttering of inBloom to the success of the Predictive Analytics Reporting (PAR) effort. PAR initially started out as a nonprofit organization, and began as a cooperative of higher education institutions that came together to create a common data framework consisting of common data elements. PAR began working with many institutions of higher education (IHEs) facing accountability pressures, including a need to report on and improve student retention and completion rates—especially for students receiving U.S. federal government dollars in the form of student loans. PAR worked bottom up with many IHEs to identify their programs designed to enhance student completion rates; to collect, clean, and organize data; and to identify the most important variables to capture. As of December 2015, PAR was using 77 variables, such as academic cycle, instructor and learner characteristics, and progress in a course, for each of an IHE's online, blended, and face-to-face courses.

How did PAR succeed where inBloom failed? Its developers and promoters would argue that, unlike inBloom, they responded to a demand that was already present in higher education. The institutions participating in PAR wanted to use data to inform improvement of their offerings and to increase their success with students.

The educational data science community can learn lessons from past failures and controversies around data privacy and security in the public and commercial sectors. The machine learning community has begun to sound its own alarms (e.g., O'Neill, 2016) and to seek solutions by developing better algorithms and better methods for using and validating models.

Bias and lack of fairness can emerge in datasets in a number of different ways. First, the datasets may not represent the full population. Withholding of student or parental consent for research participation on a large scale can mean that the data that are collected come from unrepresentative samples. Or the group participating in a digital learning experience that produces data may be self-selecting in ways that limit the generalizability of findings based on those data. In the early days of MOOCs, for example, the expectation was that populations typically without access to higher education would flock to the open courses; however, surveys showed that college-educated adults were the most likely to participate (Ho et al., 2014). Finally, because machine learning algorithms are designed to pick up patterns in data, they can pick up and reflect back the biases of the activity systems that produced those data.

One issue is the fact that big data algorithms are tuned to the majority in the dataset. Classifiers improve their accuracy if they see lots of labeled examples, but if some of those labeled examples represent a small but systematically different group, the classification algorithm may be accurate on average but not for the systematically different group. For example, when the social media site Facebook attempted to build a "fake profile" (i.e., not genuine users) detector, the profiles of Native Americans, among other minority groups, were falsely identified as "fake" (Higginbotham, 2016). A similar concern arises when a relevant feature or feature set in the data is correlated with an attribute such as race or gender: Even when not explicitly modeled, race and gender may surface as factors because they are associated with other attributes that are explicitly modeled. Thus, these factors may be correctly learned by the machine learning algorithm as predictors of outcomes, although not labeled as such. The result is that the algorithm, upon inspection of its results, appears to be making decisions based on, for example, race or gender. The machine learning expert Moritz Hardt points out that there is currently no principled way, a priori, to determine if such decisions are acceptable or in which cases they may cause harm (Hardt, September 2014). As these personal characteristics are not something an educational improvement effort can change, algorithms that point to demographic characteristics as predictors of poor education outcomes can promote a sense of helplessness among educators. These are cases where computer scientists and legal experts are working together to develop guidelines and standards for ethical practice.

Machine learning researchers are now suggesting audits on algorithms, but even with humans reviewing the models, there can be confirmation bias, based on sometimes unconscious beliefs that a developer has about social phenomena. The computing professionals' organization, the Association for Computing Machinery (ACM) issued a January 2017 statement on algorithmic transparency and accountability, citing evidence that it can be impossible to determine when algorithms produce biased or erroneous outputs. They cite three factors that may make computational models opaque: (1) the code may not be easy to explain; (2) it may cost money or reveal trade secrets to explain the code; and (3) showing input may disclose personal information. They argue that algorithmic decision making should be held to the same standards and audits as human decision making. They recommend that regulators enforce access and redress for affected groups and hold institutions accountable for decisions made by their algorithms. Developers, as part of their code of ethics, should maintain awareness of possible bias and should track data provenance, document design, and coding decisions so audits can be made, as well as rigorously validating and testing models. Researchers, knowing the susceptibility of both datasets and algorithms to various kinds of bias, should practice transparency around algorithms so that other researchers and stakeholders can replicate analyses and test their hypotheses about possible biases.

## Data Safeguards in the Age of Big Data

It is clear from legislation that has been introduced in recent years that the public's data privacy concerns are focused largely on private companies collecting and using data for purposes that are not in children's best interest. Every year since 2008, legislation has been introduced to revise FERPA, based on the perception that it does not do enough to protect student data.

At the same time, there are also voices watching out to ensure that rules do not inhibit the ability of local and state education agencies to use student data to improve teaching and learning, including conducting research with this ultimate goal. The Data Quality Campaign (DQC) is a nonprofit organization that provides information and resources, including testimony to Congress, to help administrators and legislators understand FERPA and the potential impacts of proposed changes. Other organizations that are supporting schools in data use are the Consortium for School Networking (COSN), a membership organization that advocates for effective use of educational technology in K–12 schools including broadband and wireless networking. COSN, DQC, and other stakeholders have created the Student Data Principles, which describe responsibilities and values for student data use that are relevant for all stakeholders. Other thought leaders in this space come from the Data & Society research institute, Harvard's Berkman Center for Internet & Society's Student Privacy Initiative, and the National Conference of State Legislatures.

## Conclusion

In this chapter, we unpacked the topic of privacy in general, paying particular attention to student privacy. We outlined the special concerns we have for our own data and for children's data in educational contexts and described issues surrounding ethical and responsible use of data when a dispassionate algorithm may be weighing in on consequential actions. We described discussions taking place in the computer science community on fairness, accuracy, and transparency in machine learning algorithms.

Researchers and schools engaging in data-intensive research can contribute to the public debate by documenting and disseminating their research efforts and explaining their procedures for protecting student privacy as they conduct their work. In Chapter 7 we will discuss the importance of building trust early on when we establish research partnerships with education practitioners. Trust is key to our change process, and few things can destroy trust more rapidly than sloppy practices and incomplete policies around privacy and ethical use of data.

## References

Buchanan, E. (2015). Privacy, Security, and Ethics, In Dede, C. J. (Ed.), *Data-Intensive Research in Education: Current Work and Next Steps*. Computing Research Association. Retrieved from http://cra.org/wp-content/uploads/2015/10/CRAEducationReport2015.pdf. pp. 89–92

Bulger, M., McCormick, P., & Pitcan, M. (2017). The legacy of inBloom. Retrieved from https://datasociety.net/pubs/ecl/InBloom_feb_2017.pdf

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., Seaton, D. T., & Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM*, *57*(9), 56–63. doi:10.1145/2643132

DataBreaches.net. (2015, December 12). Personal and sensitive data of 59,000 charter school students in California leaked. Retrieved April 19, 2017, from www.databreaches.net/personal-and-sensitive-data-of-59000-charter-school-students-in-california-leaked-researcher/

Data Quality Campaign. (2015). Making FERPA as simple as Green, Yellow, or Red. Retrieved April 19, 2017, from http://dataqualitycampaign.org/resource/stoplight-student-data-use/

Data Quality Campaign. (2016). Retrieved from http://dataqualitycampaign.org/resource/2016-student-data-privacy-legislation/

Har Carmel, Y. (2016). Regulating 'big data education' in Europe: Lessons learned from the U.S. *Internet Policy Review*, *5*(1). doi:10.14763/2016.1.402

Hardt, M. (2014, September 26). *How big data is unfair: Understanding sources of unfairness in data driven decision making*. Retrieved June 11, 2016, from https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

Harvard Business Review Staff. (2014, November). With big data comes big responsibility. *Harvard Business Review*. Retrieved from https://hbr.org/2014/11/with-big-data-comes-big-responsibility

Herold, B. (2014, January 22). Danger posed by student-data breaches prompts action. *Education Week*. Retrieved from www.edweek.org/ew/articles/2014/01/22/18dataharm_ep.h33.html

Herold, B. (2014, March 26). Google under fire for data-mining student email messages. *Education Week*. Retrieved from www.edweek.org/ew/articles/2014/03/13/26google.h33.html

Higginbotham, S. (2016, April 13). *Inside Facebook's biggest artificial intelligence project ever*. Retrieved from http://fortune.com/facebook-machine-learning/

Ho, A. D., Reich, J., Nesterko, S. O., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). *HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263

Leichty, R., & Leong, B. (2015). *De-identification and student data*. The Future of Privacy Forum. Retrieved from https://fpf.org/wp-content/uploads/FPF-DeID-FINAL-7242015jp.pdf

Molnar, M. (2014, December 10). Millions of Student Records Sold in Bankruptcy Case. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2014/12/10/millions-of-student-records-sold-in-bankruptcy.html

National Science and Technology Council (NSTC). (2016). *National privacy research strategy*. Washington, DC: Networking and Information Technology

Research and Development Program. Retrieved from www.nitrd.gov/PUBS/NationalPrivacyResearchStrategy.pdf

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown.

Rainie, L. (2016, January 14). How Americans balance privacy concerns with sharing personal information: 5 key findings. Retrieved from www.pewresearch.org/fact-tank/2016/01/14/key-findings-privacy-information-sharing/

Rainie, L., & Duggan, M. (2016). *Privacy and information sharing*. Pew Research Center. Retrieved from www.pewinternet.org/2016/01/14/privacy-and-information-sharing/

Stein, P. (2016, February 11). D.C. accidentally uploads private data of 12,000 students. Retrieved April 19, 2017, from www.washingtonpost.com/local/education/dc-accidentally-uploads-private-information-of-12000-students/2016/02/11/7618c698-d0ff-11e5-abc9-ea152f0b9561_story.html

U.S. Department of Education: Privacy Technical Assistance Center. (2012a). *PTAC frequently asked questions—cloud computing*. U.S. Department of Education. Retrieved from http://ptac.ed.gov/sites/default/files/cloud-computing.pdf

U.S. Department of Education: Privacy Technical Assistance Center. (2012b). *FERPA exceptions—summary*. U.S. Department of Education. Retrieved from http://ptac.ed.gov/sites/default/files/FERPA%20Exceptions_HANDOUT_horizontal_0.pdf

U.S. Department of Education: Privacy Technical Assistance Center. (2014). *Protecting student privacy while using online educational services: Requirements and best practices*. U.S. Department of Education. Retrieved from https://tech.ed.gov/wp-content/uploads/2014/09/Student-Privacy-and-Online-Educational-Services-February-2014.pdf

U.S. Department of Health and Human Services. (2009, June 23). Federal policy for the protection of human subjects ('Common rule'). Retrieved April 19, 2017, from www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html

Wallach, H. (2014, December 19). Introduction: Big data, machine learning, and the social sciences: Fairness, accountability, and transparency. Retrieved from https://medium.com/@hannawallach/big-data-machine-learning-and-the-social-sciences-927a8e20460d

Chapter 5

# Foundations of Collaborative Applications of Educational Data Mining and Learning Analytics

In this chapter, we discuss three major influences on the collaborative data-intensive improvement (CDI) model introduced in Chapter 1—data-driven decision making, emerging models of how researchers and education practitioners can engage in collaborative research, and the rapidly expanding fields of educational data mining and learning analytics. In discussing these major influences, we draw historical connections and highlight the ways in which research on data use in schools—and the factors contributing to increased data use—can be used to better understand the present and potential future for data-intensive research in education.

## Influence #1: Data-Driven Decision Making in Education

Interest in using data to improve education and learning outcomes did not originate with educational data mining or learning analytics. To the contrary, there have been efforts to promote data-driven decision making in education since at least the mid-1980s (Popham, Cruse, Ranking, Sandifer, & Williams, 1985; Sallis, 2005). Many of these efforts were inspired by Total Quality Management and continuous improvement initiatives in the business sector (e.g., Schmoker, 1996) with their origins in the work of W. Edwards Deming and other productivity experts (Sallis, 2005).

Deming's ideas about organizing industry to produce high-quality products were grounded in his experience working at Western Electric's Hawthorne plant —home to efficiency research that gave rise to the concept of the "Hawthorne effect" (Gillespie, 1991)—during the 1930s and his later work with statisticians at the U.S. Department of Agriculture. At the Western Electric plant, Deming studied work groups and became convinced that leadership style and group norms were key elements to productivity. At the Department of Agriculture, he and Shewhart, a fellow statistician, developed the idea that iterative "Plan, Do, Check (or Study), Act" cycles could be used to reduce waste and delays in manufacturing (Sallis, 2005). This idea was core to what became known as Statistical

Process Control, a set of techniques that piqued the interest of Japanese industrialists trying to rebuild their economy after World War II, and Deming went on to work with Japanese clients to develop the concepts, processes, and tools for Total Quality Management (TQM). From their origins in the manufacturing sector, TQM concepts and approaches later spread to service and financial industries (Sallis, 2005).

Concern on the part of U.S. businesses and policymakers about their ability to compete with Japan and other countries in the more competitive global market place of the 1980s fueled American businesses' interest in these techniques for promoting more consistent quality and productivity. In 1987 Congress passed the Malcolm Baldrige National Quality Improvement Act, creating a National Quality Award for performance excellence administered by the Department of Commerce. Still in effect today, the Baldrige Award competition uses a quality framework very consistent with Deming's work that includes the category of excellence in "measurement, analysis, and knowledge management."

At the same time, public-private partnerships between businesses and education entities were becoming increasingly widespread, and it was probably inevitable that TQM would move into the education sector. In 1991, an education-specific version of the Baldrige Quality framework was developed, and in 2001 the Chugach School District in Alaska became the first education organization to win a Baldrige award. Books such as Victoria Bernhardt's *Data Analysis for Continuous School Improvement*, which is now in its third edition, sought to guide education administrators in the application of TQM processes and principles to school systems.

Important principles from TQM that apply to education as well as other sectors include:

- Quality improvement is a way of working, not a single event or project.
- Continuous improvement efforts require teams that transcend organizational boundaries and where ideas can bubble up from any level of the organization.
- Organizations need to measure process as well as inputs and outputs.
- Measures need to be developed by and valued by members of the work team rather than imposed by others.

An example of a well-resourced implementation of data-driven decision making in a school district is described in Box 5.1.

### *The Impact of No Child Left Behind*

Advocacy on the part of business partners and consultants persuaded many education leaders that education decisions should be based on objective data rather than instinct or philosophy. Many more educators, however, were pushed toward data use by the requirements imposed by the 2001 reauthorization of the Elementary and Secondary Education Act called No Child Left Behind (NCLB). To receive federal funding for education, which almost all public school districts do, NCLB required extensive testing of students. Local education agencies had to test every student every year in grades 3 through 8 and one year during high school in reading/language arts and in mathematics. Further, the act required reporting results of this testing not only for students overall, but also for student subgroups defined by race/ethnicity, poverty status, special education status, and English Learner status. Finally, the act stipulated that each state had to make "adequate yearly progress" toward bringing all students in all of these student subgroups to proficiency on these examinations by the 2013–14 school year; it was up to each state to set its own definition of what constituted "proficiency." Schools that did not meet what their states defined as adequate yearly progress on improving the proportion of proficient students in each subgroup for two successive years were subject to a set of sanctions, including the requirement to let their students transfer to a better-performing school, the requirement to offer free tutoring, reduction in federal funds, and even school closure.

## Box 5.1 An Example of Data-Driven Decision Making in Education

A school district where we interviewed both teachers and administrators described its determination to increase its students' scores on state achievement tests and to reduce performance gaps between white and African American students. With these goals in mind, the district had invested in a new software system that would give not only principals but also individual teachers access to test score data for the set of students for whom they were responsible. After noticing that most teachers at some district schools were accessing the student test data but teachers at other schools weren't, the district decided to implement a district-wide initiative to form professional learning communities (PLCs) to promote the use of data. Formal PLC meetings conducted during the regular school day became district policy. There were PLC meetings dedicated to language arts and math staff development during which reading or math specialists planned and led the meetings to discuss student data and instructional strategies that teachers could use with students who had not attained proficiency in these subject areas. Other PLC meetings were led by team members to monitor student improvement, develop common assessments, and share best practices. The district continued to keep track of teachers' use of the data system. A log was kept of teachers' interactions with the data system, and principals talked to teachers who were not using the system to find out why.

When the district introduced mid-year benchmark assessments, teachers found they had more incentive to use the data system because it could offer them more recent and more detailed data on student performance. Teachers lauded the fact that they could see how individual students performed on particular test items. They were also able to compare their own students' performance with performance in other classes, the school as a whole, and the district as a whole. Because they could compare their class with other classes in their own school, teachers could have some basis for inferring whether their students' performance on particular items was due to their instructional practices or to the wording of the test item, which would affect students in all classes. This helped teachers decide whether they needed to reteach the topic related to items their students stumbled on or to make changes to their own teaching practices. Some teacher PLCs began using their common planning time to compare performance of their respective classes on the common test and to discuss the relationship between different instructional approaches and student performance. One such example came from a pair of social studies teachers who taught the same content but had students with quite different performances on the benchmark test. After the meeting, the teacher whose students had performed better began coaching his colleague.

Some teachers have found it useful to share results of interim assessments not just with colleagues but with their students. One teacher noted, "The kids even like to look at it [the interim assessment report]. That's so cool. We talk about how we need to analyze the question."

Source: Adapted from case study descriptions in Means, Padilla, and Gallagher (2010).

Not surprisingly, educational administrators became very motivated to collect and manage data that would help them assess and improve their schools' status with respect to NCLB-related requirements for improving achievement (Marsh, Pane, & Hamilton, 2006). This need helped drive district development of data warehouses and acquisition of more sophisticated student information systems, as described in Chapter 3. Further, NCLB stipulated that teachers should use data from standardized, state, and national assessments in their instructional decision making, thus linking data-driven decision making to the use of different forms of instruction for different students (Dunn, Airola, Lo, & Garrison, 2013). Further motivation was provided by the requirement in the No Child

Left Behind legislation that schools receiving federal education funding engage in "evidence-based practice." This same push for the use of practices and products with evidence of effectiveness was carried through in the grants programs administered by the U.S. Department of Education.

Educational applications of data-informed quality improvement efforts during the NCLB era typically used data on student demographics (i.e., membership in any of the subgroups for which NCLB reporting was required) and test scores—either end-of-year state tests or district-selected progress (i.e., benchmark) tests designed to predict how well students would do on the end-of-year state exam (Marsh et al., 2006). Thus, these analyses emphasized inputs (i.e., students' characteristics and which teachers provided their instruction) and outputs (i.e., examination scores), but paid scant attention to processes (i.e., what actually happened in classrooms).

Researchers who studied the use of data in schools and districts during this period uncovered a number of issues that undermined the effectiveness of this form of data-driven decision making as an improvement strategy (Means et al., 2010). First, there was the problem of the timeliness of the data. The state test scores required for federal reporting and computing adequate yearly progress were not available until as much as six months after the spring end-of-year testing. This meant that by the time the scores were available to districts, principals, and teachers, the students who had earned those scores had moved to the next grade. The scores could tell a teacher how her students fared the prior year, but she was no longer responsible for working with those particular students. What's more, classroom instruction requires literally thousands of decisions about what to say and do every single day. The time frames of annual testing and instructional decision making were badly out of sync (Crawford, Schlager, Penuel, & Toyama, 2008).

In addition, the information available from state tests is generally at a pretty coarse grain size. State test composite scores for a subject such as mathematics are generally quite reliable. Scores for more specific skill or knowledge areas within the subject, such as being able to solve measurement or fractions problems, on the other hand, are often not reliable enough to support making decisions about what to do with individual students because relatively few test items comprise the specific scales. Adding enough items to make each component scale reliable would mean a major increase in the amount of time devoted to state testing, an option that has little appeal. But when the only reliable score you have for a student is the composite subject score, all a teacher knows about that individual student is that he had a high, medium, or low score in mathematics the prior year. Such scores do not provide the kind of specific, detailed information about where the student did well and where he did poorly that could be used to plan an instructional program tailored to the student's needs.

To some extent, educators have addressed these two challenges—obtaining more recent and sufficiently detailed data on student competencies—by using formal assessments more frequently, administering them at multiple points throughout the school year, which are variously called "benchmark," "interim," "formative," or "progress" assessments. These tests can be administered at whatever interval the district or school chooses, and they can provide quite detailed results for current students since the entire year's worth of content does not have to be covered on any one of the assessments. Use of interim assessments to guide instruction became a common practice in the 2000s, with many commercial vendors and individual school districts developing assessment instruments for this purpose (Marsh et al., 2006).

Use of interim assessments could not address another important challenge to data-driven decision making based on test scores, however—the fact that scores on state tests can be improved significantly without really addressing deeper learning. Coaching students on how to work with different item formats and distributing instructional time across different content areas in a way that matches their relative representation on the state test were found to improve students' test scores (Schmoker, 1996). Visiting one school identified as exemplary in data-driven decision making, for example, we learned that teachers attributed their year-to-year improvement in students' reading scores to the fact that they had switched from an emphasis on teaching reading comprehension to one on learning new words after their analysis of the state test revealed that it had more items focused on vocabulary than items requiring understanding of passages one has read. As such "teaching to the test" caught on, scores on state high-stakes tests rose, but student achievement levels as measured by the National Assessment of Educational Progress, which is not used for accountability purposes, did not (Koretz, 2008). Teachers were learning to tune their instruction to the characteristics of the test used in their state, without generating broader learning that would show up on other measures of student achievement.

Additional challenges for NCLB-era data-driven decision making included the sometimes cumbersome data dashboards teachers were expected to work with and inadequate professional development on how to interpret student data and what instructional choices to make in response to it (Mandinach & Gummer, 2016; Means et al., 2010).

## Influence #2: New Forms of Collaborative Education Research

As described in Chapter 1, our approach to data-intensive improvement comes out of a broader tradition including multiple conceptions of education research linking research and practice. The next portion of this chapter describes some of the major collaborative education research approaches and the insights we have gleaned from them. These approaches may be thought of as a family of similar stances toward education research, and indeed specific research projects may be difficult to classify as falling into one or another of these categories. All of these approaches involve researchers and practitioners working together to learn from data, but their central purposes and collaborative structures vary, as described later.

### *Translational Research*

Within the field of medicine, there is now widespread recognition that research findings from controlled experiments in laboratory settings are insufficient to guide the actual practice of medicine. Out in the real world, where patients may not disclose all of their symptoms or may fail to follow doctor's instructions, treatments that work under controlled settings may prove ineffective. *Translational research* attempts to bridge the gap "from [scientists'] bench to bedside." By 2008 the National Institutes of Health in the U.S. and the parallel agency in the U.K. had invested over $1 billion in translational research centers for health science (Brabeck, 2008).

Educators and education researchers point out that the gap between experimental laboratory research in learning and what happens in schools and classrooms is easily as large as the gap faced in public health (Brabeck, 2008). To actually improve education outcomes, we need more than the fundamental principles of human learning as discovered and demonstrated in controlled laboratory tasks. We need to understand how multiple learning, social, and emotional processes play out in the context of academic tasks as they occur in complex classroom environments (National Research Council, forthcoming).

Daniel (2012) points out that some academic researchers have had a propensity to extrapolate recommendations for education practice and even whole interventions based on their laboratory findings without first establishing their usability and potential unintended side effects in real-world education settings. Sometimes referred to as *implementation research*, translational research that tests out practices designed on the basis of laboratory findings in classrooms and schools is time-consuming and difficult, yet vitally important.

Daniel describes translational research connecting learning science and education in terms of five steps or stages:

- Identification of findings from controlled laboratory studies with implications for classroom practice;
- Replication of the basic laboratory findings in classroom contexts to yield a "promising principle";

- Design and development of a method incorporating the principle that can be carried out by teachers in classroom settings;
- Experimentation on the promising principle in a wider range of representative educational settings to establish a "promising practice"; and
- Continued refinement and dissemination of the practice as it scales.

Key to translational research is the involvement of people with deep understanding of the contexts to which laboratory findings will be translated in the design and development of methods for classroom application.

Executing all of these steps can take a decade or more of work (Roschelle et al., 2010). It should be noted also that the different steps in translational research require different skills and knowledge bases. Many of those who study basic learning principles in university laboratories have limited acquaintance with the conditions under which public school teachers work and with current education policies and priorities. A learning principle that helps foster retention when subjects in a laboratory experiment try to memorize arbitrary word pairs may be swamped by other influences when sixth graders from disparate cultural backgrounds try to learn the order of operations in solving equations. A format that works well for college student subjects spending an hour participating in a laboratory study may try the patience and attention spans of students receiving an entire course presented in that format. Laboratory-demonstrated principles like spaced practice (i.e., learning facts and procedures is facilitated if practice is spaced out over time rather than done all at once) can run afoul of district pacing charts that specify what a class should be doing each week of the year.

Teachers need more than general principles; they are looking for specific guidance about what they should do when. Translational research teams need designers who are steeped in the culture and conditions of the kinds of classrooms where the research-based practice will be implemented. These designers can help researchers develop interventions that are compatible with classroom conditions as well as encapsulating basic learning science principles.

Translational research ideas have had an influence on the way that federal agencies think about different types of educational research (e.g., the *Common Guidelines for Education Research and Development* by the Institute of Education Sciences and National Science Foundation, 2013). The Common Guidelines developed by these agencies describe six types of research falling into three broad categories:

- Foundational and early stage research which encompasses both basic research testing theories of learning and exploratory research examining relationships among constructs that are important in learning theories;

- Design and development research to develop interventions based on foundational and early stage research and try them out in realistic settings to make sure they can be implemented and appear likely to produce the intended outcomes; and
- Impact studies testing whether the intervention actually causes the desired outcomes to occur at least under ideal circumstances, and later under a wider range of conditions.

Collaborative data-intensive improvement starts out in the second of these research categories and then moves to the third. The nature of the work in the third research category tends to be rather different from that described by the federal agencies, however. The focus for education practitioners will remain on their own jurisdiction, and they will strive to provide the ideal circumstances for the intervention to have the desired impacts in all their schools and classrooms.

Key tenets from translational research that have influenced our own thinking are the importance of:

- involving users (e.g., teachers and students) in designing practices or materials that leverage findings from the research lab;
- studying the effectiveness of practices in multiple real-world settings as implemented by the people in those settings; and
- employing rigorous research methods with appropriate control groups to estimate impacts of the new practice.

### *Design-Based Implementation Research*

When education researchers study educational practices and programs in situ (i.e., in actual schools and classrooms), they inevitably are confronted with the variability of impacts across settings. An approach that appeared to work well in one class or school often has no effect and sometimes is even detrimental elsewhere (Cronbach & Snow, 1977; Means & Harris, 2013). Such variability in findings makes it difficult for researchers to make unconditional statements about "what works" and can discourage educators from even bothering to look at research findings.

A significant body of research on the implementation of complex education innovations has emerged over the last several decades, often in conjunction with large-scale studies of an intervention's effectiveness. For example, in conjunction with their experimental studies of the impacts of innovative middle school mathematics curriculum units, Jeremy Roschelle and colleagues (e.g., Roschelle & Shechtman, 2013) investigated the extent to which effectiveness varied for different kinds of students, different kinds of teachers, or different teacher practices. Such implementation research reflects the complexity of education systems and brings issues of context and variability to the fore. But this kind of implementation research is still research done *to* and *for* educators rather than *with* them.

More recently, design-based implementation research (DBIR) has emerged as a set of principles consistent with real partnerships between researchers and practitioners. Penuel, Fishman, Cheng, and Sabelli (2011) made the first formal statement of principles for this approach in a widely disseminated *Education Researcher* article. They posited four defining principles: DBIR

- Focuses on problems of practice as conceived by multiple stakeholders.
- Requires a commitment to collaborative, iterative design processes.
- Seeks to advance theory as well as classroom learning and knowledge of implementation issues.
- Develops organizational capacity for sustaining system change.

Arguably the most radical departure from prior research practice is DBIR's commitment to jointly negotiating the research agenda with the practitioners who are partnering with researchers. Education research as usual has the researcher formulating a research question about a particular practice or program and then recruiting education entities willing to implement it as defined in the research protocol. In DBIR, the researcher first forms a partnership with education entities based on a shared general concern and *then* negotiates the research questions with them.

DBIR also borrows notions of co-design, with the idea that the best way to develop practices and materials that are really usable in schools and classrooms is to design them with members of the groups that will be using them—whether administrators, teachers, or students. Penuel et al. (2011) point out that successful scaling of an educational innovation depends on the actions of local administrators and teachers who will interpret the innovation and adjust it for local circumstances. Working with local actors involved in this process of interpretation and adjustment provides researchers with the opportunity to contribute to beneficial impacts and to better understand the conditions that facilitate or hinder effective implementation.

The collaborative nature of DBIR calls for multiple cycles of design, implementation, and refinement with the practitioners involved in the partnership participating in design and refinement activities as well as implementation (Supovitz, 2013). In DBIR, the grounds for introducing modifications and testing them in new implementation cycles are varied. They may include prior research, developers' insights, and users' or teachers' suggestions. Controlled experiments may be run, but in contrast to translational research, DBIR research–practice collaborations believe they are required only for major changes or choices involving significant risk if the wrong option is chosen (Means & Harris, 2013).

In contrast to the basic assumption underlying lists of "effective practices" or the What Works Clearinghouse of interventions with research evidence, the assumption behind DBIR is that educational interventions are not fixed objects but sets of practices that will be adapted to local circumstances and can be expected to undergo modifications and, hopefully, improvements throughout their lifespan (Datnow, Hubbard, & Mehan, 1998; Means & Harris, 2013).

Under the DBIR model, as is true in implementation research in general, the implementation of an intervention in particular settings is itself an object of research and a critical part of understanding how to scale an intervention without diluting its effectiveness.

Elements of DBIR that are central to the approach taken in our own work are:

- Joint negotiation of the research focus and co-design of the intervention involving researchers and practitioners working together and
- Extended periods of collaboration involving multiple cycles of design, implementation, measurement, and analysis.

### *Improvement Science for Education*

In recent years, a number of education researchers have taken up the challenge of studying variations in the contexts and the ways in which education interventions (i.e., practices and programs) are implemented in order to try to understand and address this variability. For example, a new inquiry-oriented science curriculum may result in better student learning outcomes in classrooms of teachers with a strong science background and the belief that all students can learn abstract concepts but not in the classrooms of teachers without these characteristics. If we can find patterns in terms of the circumstances under which a practice produces good outcomes, we can offer valuable insights about *when* and *how* it should be used, or alternatively, factors that should be addressed before the intervention is introduced.

Influential examples of this kind of work have been provided by the Carnegie Foundation for the Advancement of Teaching, which has articulated and applied what it calls "improvement science for education" under the leadership of foundation president Tony Bryk. The foundation's work has drawn on the tradition of improvement science in the healthcare industry as practiced by the Institute for Healthcare Improvement, which in turn has drawn on ideas from TQM as well as "90-day sprints" in which a change idea is generated, tried out on a very small scale, and evaluated to determine whether or not it is worth further refinement and broader implementation.

As articulated by Bryk and colleagues (e.g., Bryk & Gomez, 2008), the core phenomenon that improvement science seeks to address is variation in outcomes. Obviously, educational outcomes are not the same for every student or for every class or school. Educational researchers have spent years examining alternative approaches to instruction, and testing whether students experiencing a new approach, on average, have better outcomes than students who do not experience it (i.e., than those students experiencing "business as usual"). Bryk and colleagues argue that finding educational approaches with better average outcomes than the status quo is not enough. As shown in Figure 5.1, even when you have an educational approach for which the mean outcome for the treatment group is superior to that for the comparison group, there will still be students who received the new approach but did not attain a good outcome. In the case illustrated by Figure 5.1, this would be students in the right-hand distribution whose scores did not exceed the proficiency standard. Improvement science recognizes that if our goal is to produce reliably positive outcomes for all students, we need to know much more about the multiple factors producing educational outcomes and when and how to intervene in different contexts.

Fundamental to the practice of improvement science is the use of three guiding questions:

- What is the specific problem we are trying to address?
- What change could be made to lessen that problem?
- How can we tell whether the change we have tried is an improvement?

Like DBIR and other Total Quality approaches, such as Six Sigma and Deliverology, improvement science in education stresses the importance

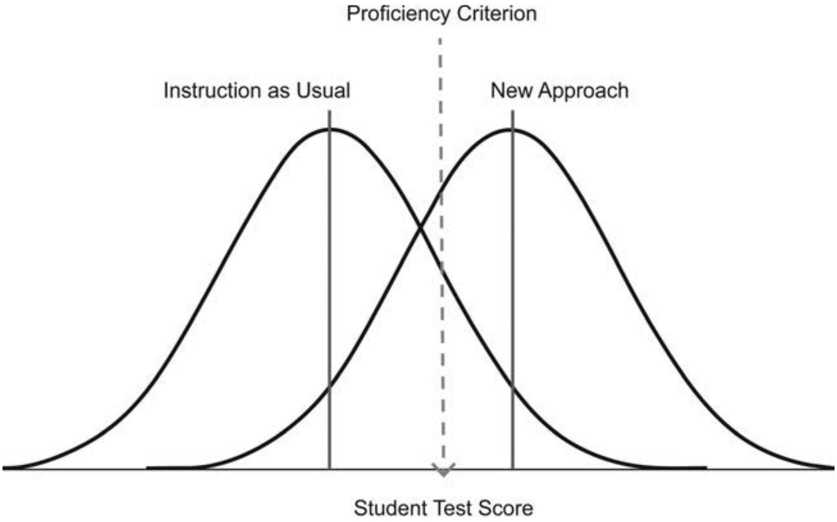


*Figure 5.1* Student Achievement Scores by Educational Treatment Condition

of multiple cycles of change implementation, measurement of processes and outcomes, analysis of those data, and planning further revisions (Bryk, Gomez, Grunow, & LeMahieu, 2015). In theory at least, each new implementation brings some degree of improvement, and the more cycles you can implement, the more positive the results.

Also similar to DBIR practitioners, those applying improvement science to education stress the importance of collaborative improvement efforts involving researchers and education practitioners working together over a sustained period of time. The Carnegie Foundation's conception of improvement science goes one step further than DBIR does, however, in highlighting the importance of having multiple educational institutions working on the same problem in parallel in a networked improvement community (Bryk, Gomez, & Grunow, 2010).

The improvement science work of the Carnegie Foundation has influenced our own practice in:

- Its focus on variation as a key phenomenon to be understood and as a source of hypotheses concerning how outcomes might be improved;
- The three overall questions that drive the research–practice collaboration;
- Its use of an articulated set of practices and standard tools to support the collaborative work of researchers and practitioners as they seek to identify and understand the problem to be studied and aspects of the education system that influence it; and
- The emphasis on "practical measures" defined in terms that make sense to practitioners and at a level of specificity that is actionable within the settings where educators work.

## Influence #3: Big Data in Education

All of the types of research–practice partnerships we have described make use of data both to identify and understand problems of practice and to be able to gauge whether new programs or changes in practice are resulting in progress toward the intended goal. But for the most part, these collaborations have not used what we would call big data or data stemming from digital learning environments, administrative data systems, or sensors and recording devices. Only in recent years have we seen very large numbers of students doing extensive portions of their learning online and thereby producing datasets ripe for data mining and analytics.

The varied uses for detailed system log data from digital learning systems have been recognized for some time. The Open Learning Initiative (OLI), for example, depicted multiple uses of micro learning data from online learning systems through the graphic shown in Figure 5.2. Founded in 2002 at Carnegie Mellon University (CMU), the OLI developed learning

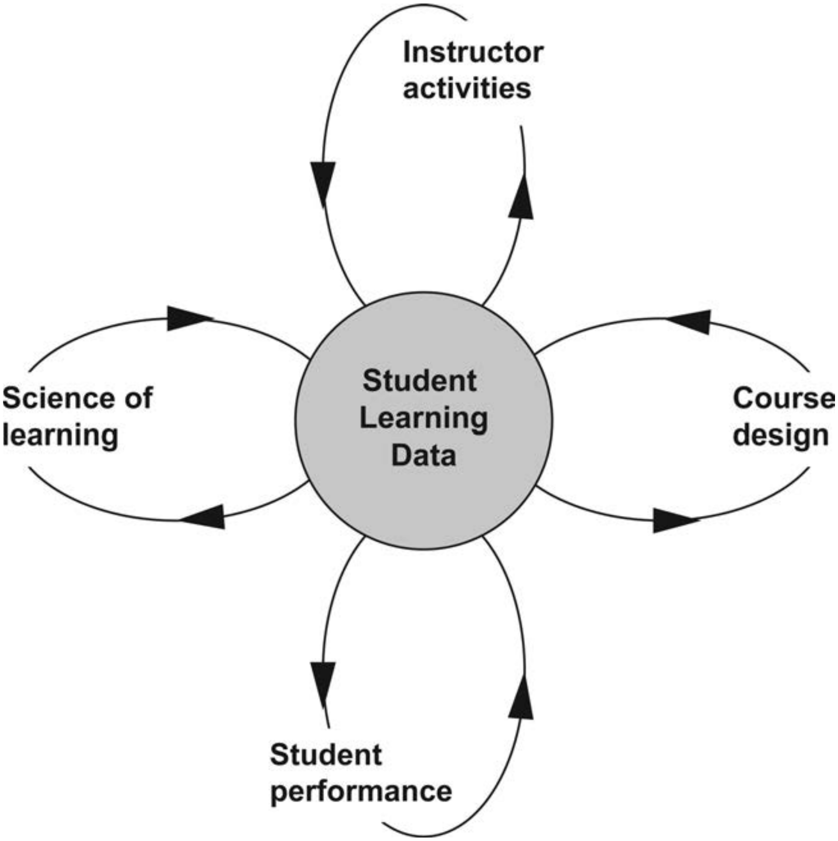
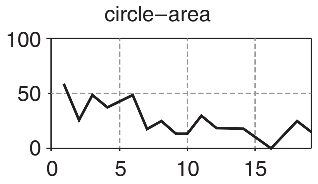


*Figure 5.2* Open Learning Initiative Use of Feedback Loops Based on Student Log File Data

systems for college courses such as Statistics, Biology, and Physics. OLI researchers explained how the data gathered automatically as students' interactions with their learning software could be useful not only for deriving measures of student performance but also for informing improvements to the learning software, providing feedback to instructors teaching a class using the software, and exploring basic questions about how people learn.

Some of the earliest efforts to capitalize on this kind of data were undertaken at Carnegie Mellon University, not only in the OLI project but also in the university's National Science Foundation-funded Pittsburgh Science of Learning Center (PSLC) and in DataShop, a repository of de-identified student-level data from multiple implementations of intelligent tutoring systems. One example of how CMU researchers use system log data to shed light on ways to improve their tutoring systems was described in Chapter 1. Here we provide a second example of how CMU researchers use detailed log file data to examine learning sequences in the form of learning curves.

Recall that CMU tutoring systems are designed using a detailed cognitive analysis of the expertise that students need to acquire. Carnegie Mellon researchers express that expertise in terms of knowledge components (KCs). One of the types of data that researchers extract from the tutoring system log files is whether the learner made an error or answered correctly on each presentation of a problem involving a given KC. Summing the number of errors made on the first trial across all learners yields an error percentage for the first trial with the KC, and the same procedure can be applied to the second and subsequent trials to develop an error curve like that shown in Figure 5.3a. Typical error or learning curves start relatively high, drop rapidly, and then level off at a low level, as illustrated in Figure 5.3a. Sometimes, however, the learning curve is much more jagged and does not trend toward zero, like the one in Figure 5.3b, suggesting that some problems related to the KC are harder than others. CMU learning scientists treat these aberrant curves as indications that their initial analysis of what is being measured by the practice items was not quite right. Sometimes, when researchers inspect all the problems mapped to a KC, they find that some of the problems appear to have some additional components not found in others.



Note: The Y-axis is the average error rate across students and the X-axis is learning opportunities.

*Figure 5.3a* Examples of Cognitive Tutor Knowledge Component Learning Curves

Source: Koedinger et al., 2013.

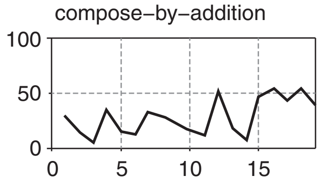
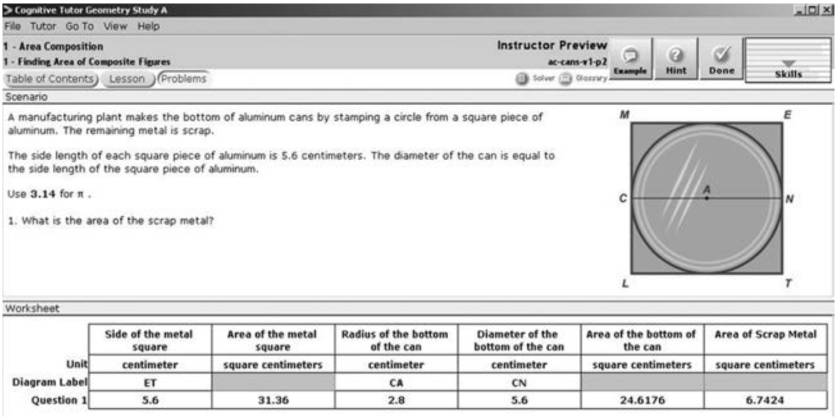


*Figure 5.3b* Examples of Cognitive Tutor Knowledge Component Learning Curves

Source: Koedinger et al., 2013.

An example comes from CMU's work with learning curves for composite area problems, like figuring out what area is left after a circle is cut from a square, in the Cognitive Geometry Tutor (see Figure 5.4a). Koedinger et al. examined the problems in the practice set for this KC and found that some of them directed the student's attention to the need to decompose the figure before trying to compute the area in question, while others did not. When working with the former, "scaffolded" items, students were not required to plan their problem solving approach. Other problems required planning in addition to area computation and subtraction. Accordingly, the researchers revised their cognitive task analysis to specify multiple KCs related to composite area problems, including one for setting the subgoals for decomposing the complex figure. The revised task analysis was then used as a basis for revising the Cognitive Geometry Tutor. The new version treated figuring out how to decompose a geometric shape into simpler shapes whose area could be computed using known formulae as a separate KC. (Figure 5.4b shows a problem designed to tap this additional KC.)

To test whether their revision of the tutoring system was indeed an improvement, Koedinger and colleagues then conducted an experiment with 96 high school students assigned to use either the new version of the tutor or the old one. They found that students using the new version of the Cognitive Geometry Tutor reached mastery on the tested skills more quickly and that they performed better on items related to problem decomposition within a paper-and-pencil post-test.



1 - Area Composition

1 - Finding Area of Composite Figures

Scenario

A manufacturing plant makes the bottom of aluminum cans by stamping a circle from a square piece of aluminum. The remaining metal is scrap.

The side length of each square piece of aluminum is 5.6 centimeters. The diameter of the can is equal to the side length of the square piece of aluminum.

Use **3.14** for π .

1. What is the area of the scrap metal?

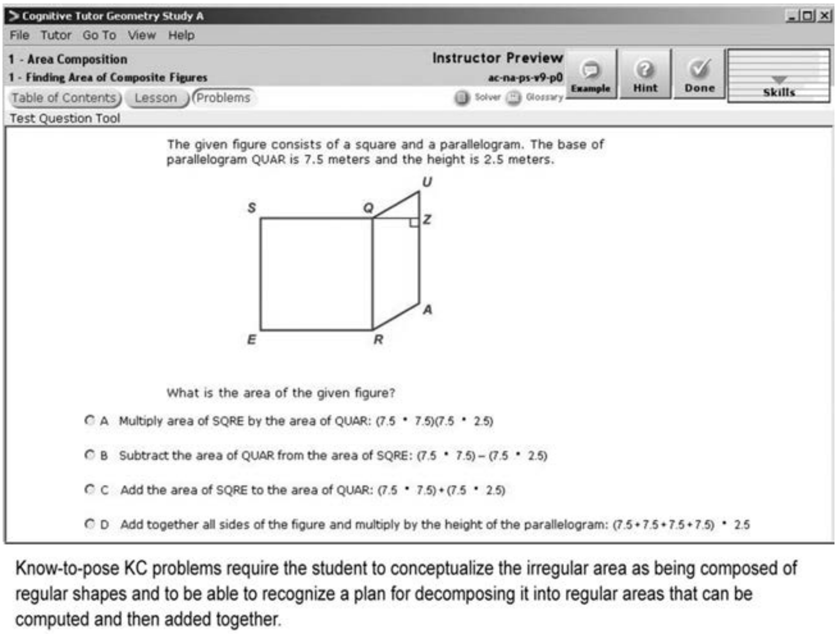
Worksheet

| | Side of the metal square | Area of the metal square | Radius of the bottom of the can | Diameter of the bottom of the can | Area of the bottom of the can | Area of Scrap Metal |
|---|---|---|---|---|---|---|
| Unit | centimeter | square centimeters | centimeter | centimeter | square centimeters | square centimeters |
| Diagram Label | ET | | CA | CN | | |
| Question 1 | 5.6 | 31.36 | 2.8 | 5.6 | 24.6176 | 6.7424 |

In these composite area problems the student fills in cell values in the table at the bottom of the screen.

*Figure 5.4a* Example of Original Version of Composite Area Problems in the Cognitive Geometry Tutor

Source: Koedinger et al., 2013.



The given figure consists of a square and a parallelogram. The base of parallelogram QUAR is 7.5 meters and the height is 2.5 meters.

What is the area of the given figure?

○ A Multiply area of SQRE by the area of QUAR: (7.5 • 7.5)(7.5 • 2.5)

○ B Subtract the area of QUAR from the area of SQRE: (7.5 • 7.5) – (7.5 • 2.5)

○ C Add the area of SQRE to the area of QUAR: (7.5 • 7.5) + (7.5 • 2.5)

○ D Add together all sides of the figure and multiply by the height of the parallelogram: (7.5 + 7.5 + 7.5 + 7.5) • 2.5

Know-to-pose KC problems require the student to conceptualize the irregular area as being composed of regular shapes and to be able to recognize a plan for decomposing it into regular areas that can be computed and then added together.

*Figure 5.4b* Example of Original Version of Composite Area Problems in the Cognitive Geometry Tutor

Source: Koedinger et al., 2013.

CMU researcher Ken Koedinger and his colleagues describe this kind of research as "closing the loop." They point out that much of the work in educational data mining has focused on developing cognitive models that enable making accurate predictions about future learning without taking the additional steps of designing interventions based on those models and demonstrating that the interventions actually improve learning (Koedinger & McLaughlin, 2016). Practitioners of data mining in many application areas are satisfied with demonstrating their ability to make predictions with a high level of accuracy, without trying to hypothesize underlying causes. Koedinger and colleagues argue that such "black box" approaches are not appropriate for educational data mining. They reason that it's important to be able to interpret predictive relationships because it is the interpretations that will lead to better understanding of learning and domain content as well as advances in instructional design. In addition, it is the interpretation that will guide the application of the predictive model to new datasets.

The Carnegie Mellon learning science research projects have benefited from having multidisciplinary teams, including faculty who were experts in the course subject matter, instructional designers, instructors using the learning software, learning science researchers, and data scientists all working together. Relatively few other learning technology R&D efforts before or since have brought together such a range of expertise. More commonly, particularly in the commercial sphere, subject matter experts and instructional designers develop learning technology products, and data scientists are brought in at a later time to identify ways to exploit user log data to identify opportunities for making the digital learning experience more appealing, with the goal of expanding the number of people who use the software for extended time periods. The primary goal for data mining in this case is similar to that in e-commerce—to increase the "stickiness" of the online experience so that people spend more time with it. Reducing even modest barriers to use (e.g., being able to do what you want with one click is much better than needing two or three) and personalizing the experience are typical goals in this work.

Other work aimed at digital learning product improvement has leveraged the existence of massive numbers of users to perform A/B tests. The Khan Academy, for example, has an A/B testing framework that enables the organization to randomly assign users to one of two or more versions of their software (i.e., version A or version B) with just a single line of code. Developers can determine what percentage of their users they want to receive each version of the software, and a dashboard charts user statistics from the two groups in real time. Box 5.2 illustrates how Khan Academy has used random assignment of subsets of their users to version A and version B of their software to find out how specific changes to their software would affect learner behavior.

In contrast, academic data science researchers have focused their efforts on developing methods that can be used with log file data from different kinds of learning systems, including the extremely large datasets generated by massive open online courses (MOOCs). In addition to describing the development and validation of computational tools and techniques, academic publications in the field focus on criteria for determining the most appropriate technique for different types of datasets. Educational data mining experts note that educational datasets differ from those typically encountered in other fields in that (1) data typically are more difficult to obtain because they must be secured from multiple organizations and in a climate of strong concerns over student privacy; (2) data on a single individual are often obtained from an extended time period, raising issues of how to combine data points into chunks that can be interpreted; and (3) data are hierarchical and non-independent as students are clustered within classrooms that are clustered within schools and districts, with each grouping exerting effects on what students do (Romero, Ventura, Pecheniziy, & Baker, 2011).

The three influences described and discussed in this chapter provided the foundation for our initial thinking about collaborative data-intensive improvement. They shaped the kinds of collaborations we have sought with education partners and sparked our interest in combining data science and improvement science methods. We will close this chapter with an extended hypothetical example designed for the purpose of illustrating the differences between CDI and its close cousins, data-driven decision making and design-based implementation research.

### Box 5.2 Combining Analytics and A/B Testing to Refine the Khan Academy Learning Platform

From its early days as an organization, Khan Academy has supported its learning technology design and implementation with Wall Street–style data analytics. In 2011 Khan Academy brought on Jace Kohlmeier, who was previously a trading systems developer at a hedge fund, as its "Dean of Data Science."

Kohlmeier designed Khan Academy's A/B testing framework to enable the organization to randomly assign users to one of two or more versions of the software with a single line of code. Developers could determine what percentage of their users they wanted to receive the experimental version, and a dashboard would chart user statistics from the two treatment groups in real time. Because Khan Academy has tens of thousands of active exercise users doing several million problems each day, developers can accrue statistically significant data very quickly. For a change with a large impact, Kohlmeier reported they can collect results in an hour (i.e., because large effects can be detected with small samples). But many of Khan Academy's experiments involve changes with smaller effects and hence take longer. In addition, the organization likes to run experiments for a week or so because of user flow cycles; more adult and self-driven learners use Khan Academy in evenings and on weekends.

One of Kohlmeier's first projects with Khan Academy was to look at how the system determined that a learner had reached proficiency on a problem set topic. The system was using a simple but arbitrary heuristic: If the user got 10 problems in a row correct, the system decided the user had mastered the topic. Kohlmeier examined the proficiency data and found that the pattern of correct/incorrect answers was important. Learners who got the first 10 problems in an exercise set correct performed differently later on than did users who needed 30–40 problems to get a streak of 10.

Kohlmeier built a predictive model based on estimating the likelihood at any point during an exercise set that the next response would be correct. The system was then changed to define mastery of a problem set as the point where a user has a 94 percent likelihood of getting the next problem correct.

This change in the mastery algorithm meant that some users had to spend more time on an exercise set. By monitoring user data after making the change, Khan Academy analysts were able to see that users were willing to devote the extra effort. At the same time, the new criterion allowed fast learners to gain credit for mastering material after doing as few as five problems, enabling them to cover more material in a given timespan. The Khan Academy team used A/B testing to compare the old and the new models for determining mastery on several metrics. They found that the new mastery model was superior in terms of number of proficiencies earned per user, number of problems required to earn those proficiencies, and number of exercise sets attempted.

In 2014 Khan Academy reworked their internal A/B testing tool (Wang, 2014). Their revised tool makes it even easier to run experiments on small interface changes intended to affect specific metrics, such as the number of problems attempted, while also supporting more complex experiments and custom analyses. To help everyone in the organization keep track of the many A/B experiments going on at any one time, the new system requires every experiment to have an "owner," a descriptive title, and an explanation of the hypothesis being tested. When analysis of experimental results is completed, the conclusions are entered into the system, and instant messages go out to alert all staff members of the new finding. Khan Academy has applied for a patent on its A/B testing tool.

Although a great proponent of A/B testing and data mining, Kohlmeier, who left Khan Academy in October 2014, notes the limitations of those approaches. It is difficult to use A/B testing to guide big changes, such as a major user interface redesign: Too many interdependent changes are involved to test each possible combination in a separate experiment.

Sources: Interview of Jace Kohlmeier conducted by Barbara Means in 2012 for the report *Expanding Evidence for Learning in a Digital Age* (Means, 2014) and interview of Alan Pierce, the software engineer who reworked the A/B testing application in 2014, conducted by Kendrick Wang (2014).

## An Illustration of Collaborative Data-Intensive Improvement

Why would a school, college, or larger education system choose to get involved with all the complexities of learning analytics and improvement processes? We offer another hypothetical example of the power of this approach relative to the alternatives. Suppose a large school district has identified achievement gaps in mathematics among middle schoolers as a key problem. District personnel reason that identifying lower-achieving eighth-grade students who do not appear to be on track to take Algebra I in grade 9 and giving them extra or different kinds of mathematics instruction could reduce disparities in the percentages of students from different subgroups who go directly into Algebra I at the start of high school. They know this is important because students who do not successfully complete Algebra I by the end of grade 9 are much more likely to drop out of high school (Silver, Saunders, & Zarate, 2008).

Let's imagine how the various traditions described in this chapter might play out in a local middle school within this district.

Suppose the school's new principal is a devotee of data-driven decision making. Her first inclination would be to gather and examine available data. First, she might want to examine the school's algebra readiness assessment scores for rising eighth graders. Accustomed to disaggregating data by gender, race/ethnicity, language learner status, and eligibility for free/reduced-price lunch eligibility, she notes that like the district's middle schools as a whole, her school has smaller proportions of its African American, Latino, English learner, and low-income students meeting the district's criterion for being algebra ready. Suppose the "gap" between African American students' readiness scores and those of white and Asian students at this school is similar in size to that found in the district as a whole, but the gap between scores for Latino and English learner students and those of whites and Asians who are native English speakers is even larger than the district average. The principal creates a set of charts illustrating the algebra readiness gaps and presents them at a staff meeting, along with a reminder that reduction of achievement gaps and getting as many students as possible through Algebra I by the end of grade 9 are district priorities.

So what to do? The principal might call in the head of the school's math department to discuss the data. The department head points out that their school has a particularly large concentration of English learners recently arrived from Central America and that many of these students struggle with the language used in the mathematics textbooks and the standardized tests. At a conference he heard about a newly released learning software product that offers middle school mathematics instruction in both English and Spanish. Students can work at their own pace and go back and forth between presentations in the two languages. The principal and the math chair decide to purchase the new software for the school's learning lab and to require all Latino eighth graders who did not test as "proficient" on the state mathematics test at the end of grade 7 to register for a "booster" math course to be spent working with the software in the lab. At the end of the year, all eighth graders are tested on the algebra readiness exam, and the principal is disappointed to find that the percentage of Latino students qualifying for Algebra I the next year has gone up slightly but not very much. She decides to abandon the learning lab strategy.

This vignette is fictitious but not at all unrealistic. Educators are using administrative data—mostly test data and information about student backgrounds—to suggest areas in need of improvement. But notice that their efforts are pretty much atheoretical. School personnel respond to pressure from the district office in choosing what problem to attend to and what data to examine. Districts are concerned with outcomes, and hold schools responsible for them, often without specifying how those outcomes are to be improved or providing supports for improving them. Looking at the outcome data motivates staff to try something new but does not really guide them in the choice of what to try. And when the new program does not immediately have an impact on the outcomes, it is hard to say why better outcomes did not emerge.

DBIR could help address this problem by bringing the perspective and skills of researchers into the picture to complement the strengths of school staff. Having committed to working with practitioners on the problem of improving Latino and English learner students' algebra readiness, for example, researchers would bring insights from prior research to bear in a more extended process of hypothesizing the sources of English learners' difficulty with pre-algebra mathematics and brainstorming potential approaches for addressing the problem and not just the symptom (i.e., test scores). The literature on the cognitive challenges posed by middle school mathematics and the additional difficulties faced by groups stereotyped as doing poorly in math and by those who are not proficient in the language of instruction is too large to summarize here. The main point is that researchers can add learning science principles and findings from the research literature to the insights practitioners have gained from firsthand experience and the data available from assessments and administrative records.

Suppose after experiencing disappointment with the learning lab course for booster math, the principal seeks out researchers from the local university to partner with for collaborative design work. A starting place is often the co-development of a theory of action. The desired end product or outcome is having a high proportion of Latino and English learning students ready for algebra at the end of grade 8. To design an intervention to improve this outcome, one needs to understand what is required to become algebra ready and to identify changes in practice that could put those elements in place. Figure 5.5 shows an example of a theory of action for algebra readiness. Note that it decomposes the desired outcome of algebra-qualifying assessment scores into the knowledge, skills,

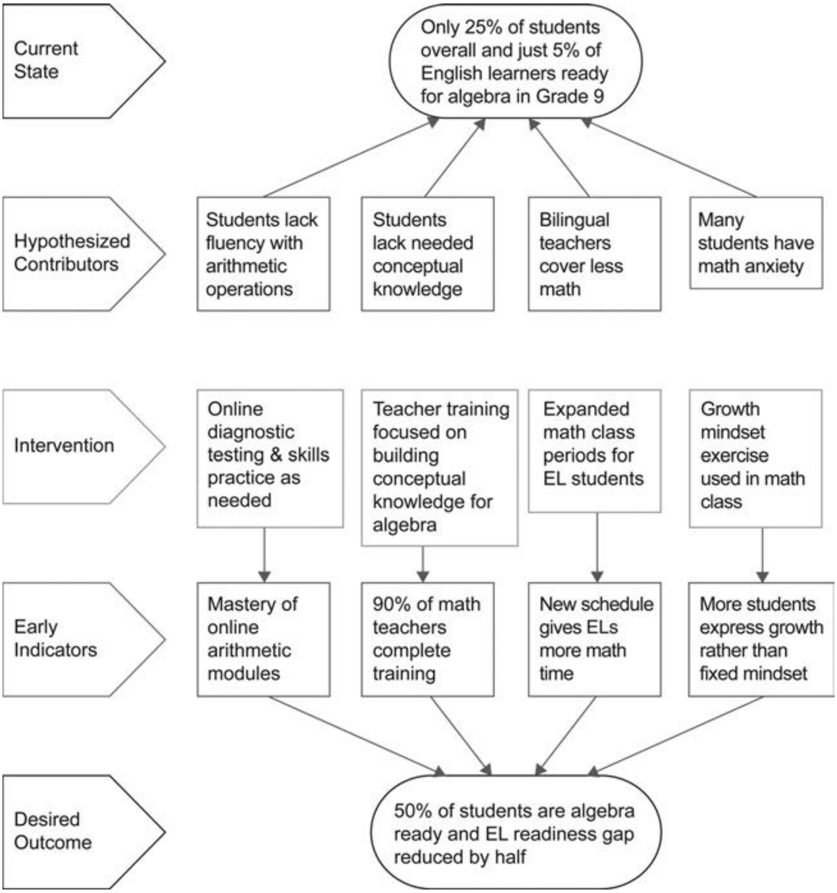

Current State
Only 25% of students overall and just 5% of English learners ready for algebra in Grade 9
Hypothesized Contributors
Students lack fluency with arithmetic operations
Students lack needed conceptual knowledge
Bilingual teachers cover less math
Many students have math anxiety
Intervention
Online diagnostic testing & skills practice as needed
Teacher training focused on building conceptual knowledge for algebra
Expanded math class periods for EL students
Growth mindset exercise used in math class
Early Indicators
Mastery of online arithmetic modules
90% of math teachers complete training
New schedule gives ELs more math time
More students express growth rather than fixed mindset
Desired Outcome
50% of students are algebra ready and EL readiness gap reduced by half

*Figure 5.5* Theory of Action for an Algebra-Readiness Intervention

and attitudes that are needed to achieve those scores. Before leaping to a conclusion about what new practices to try, a team would want to look for additional data to help identify the area in which there are widespread problems. Do students lack fluency in executing basic arithmetic procedures such as multi-digit division? Are they missing concepts such as proportion and rational numbers? Is it simply the added burden imposed when trying to read in a partially familiar language? Or are these students anxious about their mathematics performance to such an extent that it interferes with their test performance?

A DBIR collaboration around improving algebra readiness would likely deepen educators' thinking about the multiple potential sources of student difficulty and the team would then work together to test hypotheses about sources of difficulty and then to identify or co-design an intervention to address the issues that appear to hinder mathematics learning.

We see this as an important contribution of DBIR but acknowledge that compared with improvement science and TQM, design-based implementation research is a set of principles without specific tools for supporting the work. Different DBIR researchers implement the principles in their collaborations with educators in very different ways (see the chapters in Fishman, Penuel, Allen, Cheng, & Sabelli, 2013, for examples). In contrast, improvement science has a formalized set of steps and tools such as "fishbone" and "driver diagram" templates that help collaborators articulate and document their implicit theories about the main factors producing outcomes in their current and ideal systems.

Both DBIR and improvement science embrace the notion of iterative cycles of implementation, data collection and analysis, and refinement of practice. We recognize that it is much easier to declare a commitment to continuous improvement than to really put it in place. One barrier is the time frame on which educational outcome data are available. In our vignette, the principal waits until the end of the year to see scores on the district's algebra readiness assessment. If they have risen somewhat but not dramatically, the principal can modify the program and try it again the next year, but this is a very slow process given all the possible modifications that might be made. In industry, where improvement science had its origins, there were daily measures of the key outcome of flawless products produced per hour that could be tracked over time and after each innovation was introduced. Many more improvement cycles could be implemented within a given year.

To really benefit from continuous improvement processes, you need to have continuous measures of process and ways to get quick feedback concerning the impact of a change made to the system. Achievement test scores come out just once a year. Interim or benchmark tests have become popular, providing assessment results at intervals of six weeks or so, which can help address this problem of knowing early whether or not you're on the right track. But if you don't have process measures to complement these outcome measures, you still have the quandary of not knowing *why* benchmark test performance has or has not improved.

To really drive improvement, you need both (1) a well-articulated theory of the processes that produce the outcome you're looking for and the potential barriers that must be removed and (2) detailed and frequent data on the execution of those processes. CDI incorporates both of these qualities by tapping into new sources of data made available by the increasing use of technologies in schools. Borrowing techniques for working with practitioners from improvement science, CDI collaborators lay out a detailed theory of the main factors producing the current outcomes. In our vignette, these would likely be multiple, including a lack of fluency in arithmetic operations, difficulty comprehending instruction and assessment items using unfamiliar English terms, and lack of confidence about one's ability to perform well on a mathematics examination. It is likely that sources of difficulty will be different for different students, suggesting that a "one size fits all" intervention would be less than optimal. The extra period in the learning lab might be retained, but in the next iteration, the first week or so could involve building a positive class culture and diagnosing the different vulnerabilities of different students.

For students who express high levels of math anxiety or susceptibility to stereotype threat, there are simple interventions that can be done in as little as a single class period and still have significant effects (Cohen, Garcia, Apfel, & Master, 2006; Cohen, Garcia, Purdie-Vaugns, Apfel, & Brzustoski, 2009; Miyake et al., 2010). Some students might benefit from extra practice with software emphasizing executing arithmetic operations to the point where they can do them with fluency. Others might benefit from working in small groups with a bilingual teacher who focuses on concepts and has students solve problems as she watches and coaches, paying particular attention to language-based confusions. In all of these cases, a key activity will be collecting data on the execution of the strategies (i.e., How much conceptual coaching did each student receive today?) and on student performance (i.e., How quickly can the student do 20 multiplication problems with 80 percent accuracy?).

For this CDI approach to be feasible, these data must be extremely easy to collect, and having students work in a digital environment is highly advantageous in this respect. But beyond collecting data, the improvement team needs to analyze the data and make further refinements—in our vignette, each student's areas of vulnerability should be repeatedly diagnosed and his or her learning plan revised as appropriate. At the same time, educators should be looking for ways in which they can better execute the various strategies.

## Conclusion

This chapter has described the major antecedents for CDI coming from the productivity and quality improvement movement, efforts to combine educational research and practice, and data science. Our concluding vignette provided a portrayal of the kinds of information and support CDI can provide. While hypothetical, the vignette was inspired by real experiences working with and observing education institutions. In the chapters that follow, we will go beyond this fictional vignette to describe real CDI collaborations in order to illustrate the stages, benefits, and challenges of this kind of work.

## References

Brabeck, M. (2008, May 21). Why we need 'translational' research: Putting clinical findings to work in classrooms. *Education Week*, 28–36.

Bryk, A. S., & Gomez, L. M. (2008). Ruminations on reinventing an R&D capacity for educational improvement. In F. M. Hess (Ed.), *The future of educational entrepreneurship: Possibilities of school reform* (pp. 181–206). Cambridge, MA: Harvard Education Press.

Bryk, A. S., Gomez, L. M., & Grunow, A. (2010). *Getting ideas into action: Building networked improvement communities in education*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from www.carnegiefoundation.org/spotlight/webinar-bryk-gomez-building-networkedimprovement-communities-in-education

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Press.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*, 1307.

Cohen, G. L., Garcia, J., Purdie-Vaugns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science*, *324*, 400–403.

Crawford, V., Schlager, M. S., Penuel, W. R., & Toyama, Y. (2008). Supporting the art of teaching in a data-rich, high-performance learning environment. In E. B. Mandinach & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning* (pp. 109–129). New York: Teachers College Press.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington Publishers.

Daniel, D. B. (2012). Promising principles: Translating the science of learning to educational practice. *Journal of Applied Research in Memory and Cognition*, *1*, 251–153.

Datnow, A., Hubbard, L., & Mehan, H. (1998). *Educational reform implementation: A co-constructed process*. San Diego: Center for Research on Education, Diversity & Excellence. Retrieved September 2, 2016, from http://crede.berkeley.edu/pdf/rr05.pdf

Dunn, K. E., Airola, D. T., Lo, W., & Garrison, M. (2013). Becoming data-driven: Exploring teacher efficacy and concerns related to data-driven decision making. *Journal of Experimental Education*, *81*, 222–241.

Fishman, B. J., Penuel, W. R., Allen, A.-R., Cheng, B. H., & Sabelli, S. (Eds.). (2013). *Design-based implementation research: Theory, methods, and exemplars*. National Society for the Study of Education. The 112th Yearbook, Issue 2. New York: Teachers College Press.

Gillespie, G. (1991). *Manufacturing knowledge, a history of the Hawthorne experiments*. Cambridge, U.K.: Cambridge University Press.

Institute of Education Sciences and National Science Foundation. (2013). *Common guidelines for education research and development*. Washington, DC: Author.

Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., & Nixon, T. (2013). Using data-driven discovery of better student models to improve student learning. In

*Proceedings of the 16th International Conference on Artificial Intelligence in Education* (421–430). Memphis, TN.

Koedinger, K. R., & McLaughlin, E. A. (2016). Closing the loop with quantitative cognitive task analysis. Presented at the *9th International Conference on Educational Data Mining*. Raleigh, NC. Retrieved from www.educationaldatamining.org/EDM2016/proceedings/paper_152.pdf

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Mandinach, E. B., & Gummer, E. S. (2016). *Data literacy for educators: Making it count in preparation and practice*. New York: Teachers College Press.

Marsh, J., Pane, J., & Hamilton, L. (2006). *Making sense of data-driven decision making in education*. Santa Monica, CA: Rand.

Means, B. (2014). *Expanding evidence approaches for learning in a digital world*. Washington, DC: U.S. Department of Education, Office of Educational Technology.

Means, B., & Harris, C. (2013). Evidence framework for design-based implementation research. In B. J. Fishman, W. R. Penuel, A.-R. Allen, B. H. Cheng, & N. Sabelli (Eds.), *Design-based implementation research: Theory, methods, and exemplars*. National Society for the Study of Education. The 112th Yearbook, Issue 2 (pp. 350–371). New York: Teachers College Press.

Means, B., Padilla, C., & Gallagher, L. (2010). *Use of education data at the local level from accountability to instructional improvement*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development.

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, *330*, 1234–1237.

National Research Council (forthcoming). *How People Learn: Third Edition*. Washington DC: The National Academy Press.

Penuel, W. R., Fishman, B. J., Cheng, B. H., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educational Researcher*, *40*(7), 331–337. doi:10.3102/0013189X11421826

Popham, W. J., Cruse, K. L., Ranking, S. C., Sandifer, P. D., & Williams, R. L. (1985). Measure-driven instruction. *Phi Delta Kappan*, *66*, 628–634.

Romero, C., Ventura, S., Pecheniziy, M., & Baker, R. S. J. D. (Eds.). (2011). *Handbook of educational data mining*. Boca Raton, New York, London: CRC Press.

Roschelle, J., & Shechtman, N. (2013). SimCalc at scale: Three studies examine the integration of technology, curriculum, and professional development for advancing middle school mathematics. In J. Roschelle & S. Hegedus (Eds.), *Democratizing access to important mathematics through dynamic representations: Contributions and visions from the SimCalc research program* (125–143). New York: Springer.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., et al. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal*, *47*(4), 833–878.

Sallis, E. (2005). *Total quality management in education*. London: Kogan Page.

Schmoker, M. J. (1996). *Results: The key to continuous school improvement*. Alexandria, VA: Association for Supervision and Curriculum Development.

Silver, D., Saunders, M., & Zarate, E. (2008). What factors predict high school graduation in the Los Angeles Unified School District? California Dropout Research Program Report #14. University of California Santa Barbara. Retrieved from http://cdrpsb.org/dropouts/researchreport14.pdf

Supovitz, J. (2013). Situated research design and methodological choices in formative program evaluation. In B. J. Fishman, W. R. Penuel, A.-R. Allen, B. H. Cheng, & N. Sabelli (Eds.), *Design-based implementation research: Theory, methods, and exemplars*. National Society for the Study of Education. The 112th Yearbook, Issue 2 (pp. 373–399). New York: Teachers College Press.

Wang, K. (2014). How Khan Academy uses A/B testing to improve student learning. *Apptimize*. Retrieved October 1, 2016, from https://apptimize.com/blog/2014/07/how-khan-academy-uses-ab-testing-to-improve-student-learning/

# Chapter 6

# Supporting Conditions for Collaborative Data-Intensive Improvement

As the room filled up, those new to the gathering saw a modern conference room that served as both a workspace and a cafeteria. After nearly two days of work, three teams, each made up of researchers and practitioners from a Charter Management Organization (CMO) based in the Bay Area, scrambled to put the finishing touches on their presentations. This day was preceded by nearly two and half years of work. The path connecting the early ideas of a partnership between researchers and practitioners around the use of data-intensive research methods was not without twists and turns. But at the heart of the partnership, from very early on, was a sense of trust and mutual benefit that helped members of the partnership lean into the twists and turns that accompanied using new forms of data to improve teaching and learning in classrooms throughout the CMO.

The teams preparing their final presentations were formed on the first day of the two-day event, which the partners called a "data sprint." The sprint was an opportunity for the partners to come together to jointly analyze data from digital learning environments and administrative data systems. Three teams were organized around three driving questions: What is the relationship between external and internal measures of student learning and achievement? What is the relationship between students' activity in a digital learning environment and external measures of student achievement? How many distinct learning behaviors can be measured using digital learning environment data? The sprint was structured as a two-day event in order to allow sufficient time to analyze digital learning system data and co-develop follow-up actions, such as new instructional routines that teachers could later implement in their classrooms. The first of three teams began its presentation on the degree to which the CMO's own assessments correlated with external measures of students' college readiness, such as standardized test scores. Nearly every component of the presentation, as well as the accompanying analysis, had been a collaborative undertaking, from the merging of data files and transformation of variables to the interpretation of results.

At the end of the presentation, the first team laid out the next steps they would take based on what they had learned from their analyses. Using simple scatterplots with carefully chosen reference lines, they had identified a group of students who were behind on both internal and external measures of college readiness. The second and third teams, similarly, identified previously unnoticed patterns in students' use of the CMO's digital learning environment. For example, they found that students who tended to follow up a poor performance on an assessment by accessing an available learning resource did better in the course overall than students who tended to follow up a poor performance by retrying the assessment. These presentations would launch months of work examining the patterns identified during the sprint across multiple grades, content areas, and over wider timescales.

Getting to a point where researchers and practitioners were able to jointly engage in data-intensive research did not happen overnight. New knowledge, skills, and dispositions needed to be developed by both researchers and practitioners. In this chapter, we describe our approach for engaging in data-intensive research–practice partnerships (RPPs). Across multiple partnerships, we have engaged in two overarching activities: First, we collaborated with an educational organization around using data-intensive research methods to explore the organization's own data. Second, we explored ways of improving how we worked with each educational organization. These two activities served as the building blocks of a multi-year, multi-project research agenda aimed at developing a repeatable approach for collaborating with practitioners, what we have come to refer to as *collaborative data-intensive improvement* (CDI). We will return to the CMO described in the introduction of this chapter and describe how we set the foundation for the data sprint over multiple years of collaboratively analyzing the CMO's data and working to improve the partnership over time. Before returning to the CMO and introducing a second partnership, we briefly describe the origin of CDI through the lens of prior efforts to use data in schools, building on the overviews provided in the previous chapter.

## Origins of Collaborative Data-Intensive Improvement

Working with practitioners to engage in data-intensive research is not a new idea. Groups such as the Youth Data Archive (YDA) at the John W. Gardner Center for Youth and Their Communities regularly work in close collaboration with practitioners to develop research questions, interpret analyses, and co-develop changes (Russell, Jackson, Krumm, & Frank, 2013). The "data for good" movement through organizations like DataKind and Bayes Impact has demonstrated the potential for bringing together data scientists and practitioners from non-profit and non-governmental organizations. And foundations like the Ann E. Casey Foundation have funded efforts to make use of integrated data systems from non-profit and governmental service providers to better understand issues facing youth and families. Thus, there are multiple examples of researchers, data scientists, and practitioners coming together to analyze and take productive action based on analyses of large volumes of data from, in particular, administrative data systems.

As we began thinking about how to use new sources of data with practitioners, we were drawn to the work of organizations like the Harvard Strategic Data Project, DataKind, and YDA (e.g., McLaughlin & London, 2013). Across such standout organizations, though, we saw two gaps that had yet to be addressed when we started thinking about data-intensive RPPs in 2014. First, most if not all partnerships were structured around administrative data systems—we wondered what could be gained by forming partnerships around data from digital learning environments, alone or in combination with data from other sources. Second, the work involved in actually engaging in collaborative data-intensive research had not been detailed within the existing literature—we set out to develop in-depth accounts of how to engage in collaborative data-intensive research to help subsequent partnerships based on our lessons learned. We addressed these two issues by launching a series of partnerships, consulting prior research where it was available, and reflecting on what worked and what did not. Within each partnership, we engaged in a style of inquiry referred to as design-research, which we describe later, and through a series of design-research cycles we set out to identify *supporting conditions* and *key phases* for engaging in CDI.

In education, design research maps closely onto the development of the learning sciences (Bransford, Brown, & Cocking, 2000). The key insight of this approach, going to back to the pioneering work of Ann Brown (1992) and Allan Collins (1992), is that evidence collected under tightly controlled conditions, such as laboratory settings typical of early psychological research, may not in fact generalize well to the day-to-day realities of schools. This insight, which has been wrestled with by many in the field of education, developed into a methodology where researchers use theory to develop interventions that are then tested in real learning environments. Through iterative refinement and collaboration with practitioners, design research involves creating a "humble" theory for why an intervention worked (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). Many learning scientists and methodologists have weighed in on the pros and cons of design research, and it is not an approach without critics (Kelly, 2004; Shavelson, Phillips, Towne, & Feuer, 2003). Yet for some, including us, it offers a way of approaching practical educational problems while working to build theory (Bell, 2004). The key unit of analysis in traditional design research is a learning environment, such as a classroom or after-school learning experience. Through working in real-world environments, the goal is not to isolate individual causes for why an intervention worked or not but rather to attend to, as best as possible, how an intervention affects and is affected by the myriad complexities within each setting (Barab & Squire, 2004). In our work, the key units of analysis are the partnerships themselves and how the work of the partnership translates into changes in the ways that practitioners interact with learners.

At the outset of our efforts to develop data-intensive RPPs, we simultaneously embraced and questioned the assumption that data from digital learning environments would be beneficial to understanding and improving learning environments. As we noted previously, up to this point in the development of the fields of educational data mining and learning analytics, there were only a handful of collaborations between researchers and practitioners using data from digital learning environments (e.g., Krumm, Waddington, Teasley, & Lonn, 2014). Examples of collaborations built around data residing in administrative data systems were more prevalent, and a number of these involved large urban school districts developing early warning systems with university partners (e.g., Allensworth & Easton, 2005; Balfanz, Herzog, & MacIver, 2007). The paucity of meaningful collaborations around digital learning system data was unfortunate given the pitched rhetoric around "big data in education," which carried the implication that data from digital learning environments was somehow a self-activating resource. The process for translating data into new actions that teachers would take in classrooms was either ignored or assumed to be trivial (Piety, 2013). While we embraced the potential value for newly emerging types of data, we questioned how much could truly be accomplished without having educational data scientists move closer to practice and practitioners move closer to the work of educational data science.

There is a long history of attempts to bridge research and practice in education, of which design research is but one example (Lagemann, 2000). And while it is easy to say that researchers should work more closely with practitioners, and vice versa, consistently doing so has proven to be challenging. To inform efforts at bringing researchers and practitioners together, we focused on the focal activity in which researchers and practitioners would be engaged, namely, planning for, collecting, and interpreting data, commonly referred to as data-driven decision making (DDDM; see Chapter 5). Explicit models for engaging in DDDM around standardized tests and benchmark assessments existed (e.g., Boudett, City, & Murnane, 2013; Bambrick-Santoyo, 2010). However, the evidence that decisions were informed by data and that these decisions had ultimately contributed to improved instruction was far from clear (Coburn & Turner, 2011; Hamilton et al., 2009). One reason for the limited number of clear cases may have had much to do with the ultimately weak theory of change associated with most DDDM models, which involves educators—alone—developing and deploying capabilities necessary for setting goals, determining causes for success and failure, implementing new approaches, monitoring effectiveness, and reflecting on the overall DDDM process (Penuel & Shepard, 2016). As many have pointed out, there are multiple potential points of failure in this theory, and many researchers who have studied how teachers actually worked with data found that data rarely "drove" decisions because data were often difficult to access and lagged too far behind the processes that practitioners could affect (Little, 2012). Moreover, the interpretations that teachers made of the data were affected by numerous tangential factors (Turner & Coburn, 2012). The question for us, which ran headlong into assumptions that "bigger" data would lead to "better" decisions, was as follows: How would large, complex datasets and less well understood machine and statistical learning techniques interact with these underlying dynamics of data use in schools?

An essential place to start in answering this question was to bring researchers and practitioners together in a partnership that drew on the respective skills and expertise of each. It is easy to assume that schools can and should take on this work by themselves or that data scientists can overcome the challenges of data use in schools absent knowledge of schools or the active participation of practitioners, but the track record for these assumptions is not great. To understand how complex datasets and new analytical techniques could be used in schools, we identified opportunities for improvement within the existing literature and launched several partnerships organized around translating data from digital learning environments into concrete changes; and in reflecting on the multiple partnerships in which we were working, we set out to develop an overall approach for engaging in CDI.

## Two Data-Intensive Research–Practice Partnerships

We opened this chapter with a description of an experience from one of two early CDI partnerships that we launched. The CMO that participated in the data sprint and the thousands of hours of collaborative work that preceded it was Summit Public Schools. Summit is a CMO that operates schools in the Bay Area and in the state of Washington. Summit has been recognized as an innovative CMO based in no small part on its deep integration of technology and focus on providing personalized learning experiences to all its students (Murphy et al., 2014). The second partnership that we describe later involves our work with the Carnegie Foundation for the Advancement of Teaching and the Carnegie Math Pathways. As described in Chapter 5, the Carnegie Foundation for the Advancement of Teaching has become the central organization for advancing the use of improvement science in solving long-standing educational problems and inequities. As part of their field-building efforts, Carnegie launched the Carnegie Math Pathways, which is a unique network of 2- and 4-year colleges and universities focused on solving the developmental math crisis in the United States (Bryk, Gomez, & Grunow, 2010).

### *Summit Public Schools*

In 2003, Summit began as a single high school, Summit Preparatory Charter High School in Redwood City, California. Since then, Summit has grown to 11 schools and a national program referred to as *Summit Learning*. Core to the Summit model of teaching and learning is a focus on personalization and strong relationships between students and teachers combined with giving all students a rigorous, college preparatory curriculum. A typical day at a Summit school is broken up into 90-minute blocks during which students engage in project-based learning in core subject areas. Project-based learning is an instructional method where students gain knowledge, skills, and dispositions through authentic, engaging, and complex problems (Larmer, Mergendoller, & Boss, 2015). In addition to project-based learning blocks, students engage in "personalized learning time" and weekly mentoring sessions. Personalized learning time offers students an opportunity to work on core academic content at their own pace, and mentoring sessions are times when students work one-on-one with a teacher who advocates for them and helps them develop self-directed learning skills.

Every Summit student is provided with a Google Chromebook and access to a customized learning management system (LMS) referred to as the *Summit Learning Platform*. Students use the platform in all of their courses and for a majority of their overall learning activities. For example, students use the platform during personalized learning time to access required assessments and teacher-curated resources, in the form of "playlists." Completing a playlist involves passing a 10-item content assessment that students can take as many times as they need and whenever they feel ready to take the assessment. Students interact with playlists during personalized learning time, which includes two 90-minute blocks throughout the week and for extended periods of time on Fridays, which is also when students interact with their individual mentors. Through the platform, Summit students also can work on elements of projects and can communicate with their teachers about their progress on specific elements of a project. From its founding in 2003, Summit has created a learn fast culture in which all elements of the student learning experience—from mentoring to the *Summit Learning Platform*—are continuously refined over time.

Our partnership with Summit began in the summer of 2014. Prior to the start of the partnership, Andy and a senior leader at Summit had participated in a national conference on the topic of personalized learning. At the conference, Andy and the then Chief Information Officer for Summit Public Schools discussed the multiple research projects in which they had been a part. Summit was an early adopter of multiple technologies and as a CMO they had participated in multiple studies on how they used technology to support teaching and learning. Summit staff lamented the fact that researchers' insights and efforts were often directed toward writing reports, as opposed to helping Summit staff grow and improve. This observation led the two to jot down the basic outline for a partnership organized around the idea of analyzing data from Summit's digital learning environments for the purpose of improving teaching and learning at Summit. A few follow-up phone calls later, the partnership between Andy and Summit had expanded to include Alex Bowers from Teachers College, Columbia University.

In the following sections, we provide a chronological description of our partnership with Summit Public Schools. Throughout the project, the participating researchers met regularly to reflect on the partnership and to clarify lessons learned about the *supporting conditions* for engaging in collaborative data-intensive research. The goal of identifying these conditions was to help subsequent partnerships launch and organize their own work. Design-research cycles, like those described previously, were organized around key events such as initial brainstorming meetings and subsequent meetings where members of the partnership would come together to jointly analyze and interpret data products. From the start of the project, we regularly experimented with how best to bring researchers and practitioners together, and we engaged in multiple data analyses geared toward helping Summit practitioners improve learning opportunities for students.

### *Setting the Foundation*

The process for identifying the first round of research questions that would guide the partnership began with an initial, face-to-face meeting of leaders from Summit Public Schools and members of the research team. At the meeting, we engaged in a round of brainstorming activities where Summit leaders proposed topics and questions that the partnership could explore. Examples from this initial meeting included "identifying and measuring self-directed learning behaviors," "identifying the relationships between micro-momentary choices that students were making and their college-going trajectories," and "identifying specific ways to keep students on-track." Following an initial round of brainstorming, the technology and information teams from Summit outlined the data that were captured and stored by their various systems. This first meeting concluded with a preliminary set of topics for the partnership to pursue and a developing understanding of the data that could be used to explore each topic. Following the initial brainstorming meeting, the partnership blended Summit's research interests with the knowledge and expertise of the research team using the following process: (1) members of the research team wrote brief descriptions for how they could attempt to answer each question that was developed during the initial brainstorming session; (2) practitioners then reflected on the approaches proposed by the research team; and lastly, (3) the research team and Summit leaders came together to evaluate the potential impact and feasibility of answering each question.

The first question that the partnership collaborated on involved understanding patterns in students' attempting and completing content assessments. Summit's 10-item content assessments are quizzes that students are required to complete at the end of each playlist. Different courses require different numbers of content assessments to be completed. A distinctive feature of playlists is that students are given both the freedom to work on whatever playlist they choose and discretion in how they navigate each playlist. A core element of Summit's learning model is that students are provided with opportunities to grow and demonstrate "habits of success," such as self-direction, curiosity, and civic identity. Using data from the 2013–14 academic school year from the LMS used by Summit at the time, we began exploring patterns of how students took and passed content assessments, not only to answer the focused research questions but also as a concrete way to measure students' self-directed learning behaviors. Self-directed learning is closely related to self-regulated learning (e.g., Pintrich, 2004), which refers to the ways learners actively regulate their own cognition, motivation, behaviors, and elements of their environment in order to achieve a goal.

The project officially kicked off in September 2014, and we completed our first joint data interpretation meeting at the end of October. The speed with which the research team was able to complete these first analyses of students' behavior in taking content assessments was important given the partnership's goal of doing research differently and shortening the time between posing a research question and having a potential answer. What made exploring this question possible in such a short time frame was the fact that an entire prior academic year's worth of data had already been collected and stored within Summit's LMS and student information system.

To explore patterns in content assessment taking, the partnership used two initial strategies. First, we specified what content assessment taking should look like so that each student's actual content assessment taking could be compared against that normative standard. Second, the different patterns found in students' actual content assessment taking were examined in relation to outcomes that the partnership valued, such as course grades. For this first analysis, we examined student assessment taking patterns in relation to students' final grades in four core courses: Math, English, Science, and Humanities in ninth grade. For these four courses, we wrangled data from a database that tracked and stored students' content assessment taking in the LMS and from Summit's student information system, which contained students' course grades and standardized test performances.

We arrived at two major takeaways during an early fall meeting to review initial analyses. The first insight was based on the finding that the extent to which students struggled with content assessments varied for different assessments. One way in which we identified the degree to which students struggled was by examining students' scores the first time they attempted a content assessment. For some content assessments, the median student scored 4 out of 10 on his or her first attempt, while for other content assessments, the median student scored a 7 or 8. This and related findings generated questions around what was contributing to low scores (e.g., content assessment difficulty, students' prior knowledge related to the specific content being assessed, or the ways students prepared for taking the content assessment). The implication from these analyses was that effort should be directed, both by Summit practitioners following the meeting and by the research team in the form of new analyses, toward understanding factors contributing to students' struggle with particular assessments.

The second key insight from this same early fall meeting was that teachers should have easier access to cumulative and longitudinal data on students' attempts and completions of content assessments on the Summit platform. Up to this point in the history of the platform, teachers lacked basic information about what students were doing in the system. They could not answer questions such as "How many days were students taking to complete a playlist?" or "What resources were students using most often?" By aggregating a year's worth of data and structuring a conversation around how to interpret the data, the research team was able to demonstrate the potential benefit of providing longitudinal data directly to teachers through new data displays in the platform.

This initial cycle of inquiry would set expectations for the many partnership-driven analyses to follow. Direct engagement between researchers and practitioners helped improve researchers' understanding of the data they had analyzed and provided practitioners with an opportunity to see how learning at their school was playing out at scale and over time. Both kinds of insights helped the partnership brainstorm potential changes to the content assessments themselves and to how teachers worked with students to prepare for content assessments.

After this first cycle of inquiry, the research team added another dataset into the mix—students' use of the specific learning resources in their playlists. Using this additional dataset, the research team was able to examine relationships among students' standardized test performances, their use of playlist resources, and their content assessment taking and course grades. Perhaps unsurprisingly, an early finding was that students with lower incoming standardized test scores were attempting math content assessments more frequently. While lower incoming content knowledge may explain the need for more attempts in order to demonstrate mastery on an assessment, we also observed that students with higher standardized test scores were using the system in different ways than their peers with lower incoming scores. For example, higher-scoring students were using more unique learning resources and looking at those resources prior to taking a content assessment rather than afterward.

### *Building on Lessons Learned*

The second deep-dive meeting between researchers and Summit leaders was held in winter of 2014 with the goal of discussing the analytic findings regarding students' learning resource use, content assessment taking, and course performance. This meeting prompted the partnership to think about how the findings could be communicated directly to Summit teachers. Up to this point, the researchers had been working most directly with the CMO leaders, and the leaders took responsibility for communicating findings and negotiating potential changes with teachers. The partnership targeted an upcoming all-CMO professional development meeting as an opportunity for teachers to learn more about the data analysis findings and to interact directly with the visualizations and other data products that represented those findings. The partnership collectively developed a strategy for using findings from the fall and winter meetings to create datasets that could be integrated into Summit's own data management and visualization tools. The hope was that by having the researchers take care of data wrangling and giving Summit teachers the opportunity to work with the organized datasets using tools they were already familiar with, a large number of teachers could engage with these data in an in-depth way. To help teachers navigate the new, unfamiliar data elements within their familiar systems, the research team briefly presented a description of the meaning of each data element and demonstrated how teachers could use a flowchart developed by the Summit information team. The flowchart was intended to help teachers identify whether or not a playlist was ripe for revision based on how many times students attempted its content assessment and the ways in which students used the playlist's learning resources.

During the first year of the partnership, Summit practitioners and the research team engaged in three cycles of inquiry into Summit's own data based on Summit's research questions. Across multiple meetings, the partnership experimented with how to surface practitioners' research questions, how to present findings, and how to translate findings into follow-up actions. As the partnership moved into the second year, members made a concerted effort at a two-day meeting to reflect on and highlight lessons learned that would inform the second year of working together.

In reflecting on the first year, the partnership members affirmed the importance of having Summit lead the question-generation process and of having researchers support that process by reflecting on questions and the potential impacts and feasibility of addressing them. It was apparent that the speed with which researchers had answered the first question posed by Summit leaders helped build trust with Summit. Another key reflection was the way in which the partnership came to value and take seriously the fact that opportunities to meet and discuss data analyses were *learning events* as opposed to presentations where researchers would present and defend their analysis. Instead, meetings were structured as collaborative opportunities for researchers and practitioners to learn from one another. Concretely, researchers made intentional efforts to not just present findings but to make as explicit as possible the thinking that went into each analysis. For example, we included data products that were built using sample or fake data to help practitioners understand the logic behind an analysis before presenting that product using their own data. Lastly, we organized a specific kind of learning event where researchers provided training to Summit's information team on how to use the R software.

The second year of the project got under way with a return to the original 10 research questions that were generated at the partnership's initial brainstorming meeting. From the original list the partnership selected two questions to focus on: (1) How do Summit's internal metrics relate to external benchmarks for college readiness? (2) Can we characterize what students do in the platform as successful or unsuccessful? To explore how Summit's internal metrics related to external benchmarks, we analyzed relationships among students' course grades and multiple standardized test scores, using data from multiple grade levels and subject areas. We then organized a broader meeting of teachers to explore the degree to which measures collected by Summit correlated with college-readiness indicators from external organizations. Across multiple meetings, we worked hard to help practitioners understand how to interpret specific relationships between internal and external metrics. Alex Bowers, who had recently done an in-depth analysis of the relationship between grades and standardized test scores, directly supported Summit staff during this period as they interpreted the developing findings.

Following these meetings, the researchers in the partnership conducted fewer analyses of the relationships between internal grades and external test scores because Summit staff were easily managing and analyzing these data. For these types of analyses, our roles shifted to helping interpret analyses that were initiated and carried out by Summit staff.

As researchers' involvement in conducting these correlational analyses declined, we began work on the second research question concerning successful and unsuccessful behavioral patterns within the Summit platform.

While data related to student outcomes, such as course grades and standardized test scores, had not required much, if any, feature engineering, data from the platform did. Data from the platform contained millions of observations and required knowledge of learning theory as well as certain technical skills in order to turn those observations into meaningful features that could then be compared against student outcomes. Over time, the pattern of researchers' activity became clearer: When there was little need to engineer features, researchers supported school staff who conducted analyses themselves in thinking about how to interpret findings; when a lot of feature engineering was required, researchers did more of the analytic work as well as supporting data interpretation.

Characterizing successful and unsuccessful student behavior patterns using platform data required significant feature engineering. Starting with data from the then current school year, we used three general approaches to look for patterns. The first approach built off of our analyses the prior year and entailed simply summarizing how each student used the platform in terms of the number of resources used, unique resources used, resources used before taking the first content assessment, number of content assessment attempts, and similar measures of the quantity of various types of activity. The second approach involved identifying strings of events. At a general level, each playlist comprises different types of resources and assessments. We coded each digital learning event as either a resource (R) or as an assessment that was either passed (P) or failed (F). Thus, each playlist that a student worked on could be represented as a string of letters (e.g., RRFRRFP). When combined with the quantitative metrics (e.g., number of unique resources used on a playlist), the event strings characterized the many possible ways that students used the learning platform. The third analytic approach we used involved quantitatively characterizing movements from one event to another. We created metrics, using conditional probabilities, that quantified how likely it is that a student would move from a certain kind of event to another kind of event. For example, if the student has failed a content assessment, how likely is it that the student will go immediately to another assessment and fail it? How likely is it that the student will go from the failed assessment to examining a learning resource? Across each pair of event categories, we identified students' most likely transitions, i.e., their most likely next step.

We presented these different ways of characterizing students' patterns in a spring 2016 meeting with Summit leaders and teachers. Throughout the summer of 2016, we used known grade and achievement score outcomes from the 2015–16 school year, and worked to identify more and less successful learning behavior patterns. Naturally, "it depends" was a common phrase in our discussions. For example, students who came to a playlist with a high level of domain-relevant knowledge tended to use few resources and ultimately needed fewer attempts to pass the required content assessment. To the question "Should students be using more learning resources?" *It depends*. Across multiple analyses, we developed evidence for a variety of patterns that members of the partnership took with them into the data sprint described in the opening of this chapter. In bringing researchers and practitioners together for a concentrated amount of time to explore new questions, we attempted to set up specific tests of change that could be enacted following the sprint—thus moving the partnership toward a better understanding of more or less successful learning behaviors.

### Carnegie Foundation for Advancement of Teaching

During the same summer that our partnership with Summit began, Andy and colleagues started working with researchers and staff at the Carnegie Foundation for the Advancement of Teaching. At the time, Carnegie was well into launching and supporting the Carnegie Math Pathways, which is a national effort focused on improving developmental, or remedial, math courses in 2- and 4-year colleges throughout the United States. In some colleges, students can be required to take a developmental mathematics course if they have been identified as not ready for college-level mathematics. These courses can be a significant barrier to college completion; only a small proportion of the students who are required to take these courses pass them and go on to earn the college-level math credits required for many degrees. One study of 57 community colleges found that 80 percent of the students assigned to a sequence of developmental math courses did not successfully complete a transfer-level (i.e., credit-bearing) math course within three years (Bailey, Jeong, & Cho, 2010). The number of lives affected by the developmental math crisis in the United States is staggering, so Carnegie brought together experts in mathematics education with college teams who all wanted to tackle this problem using a new and promising set of approaches referred to as improvement science.

With Carnegie as the hub, they formed a networked improvement community (NIC), which as introduced in the previous chapter is a type of scientific community that is organized around a common aim, guided by a common understanding of the problem it is trying to solve, disciplined by the use of improvement science tools, and deliberately structured to share knowledge across those participating in the network (Bryk, Gomez, Grunow, & LeMahieu, 2015). Members of the Carnegie Math Pathways

NIC designed two different course sequences, or *pathways*, representing alternative, intensified approaches to fulfilling developmental math requirements and earning college credit in either statistics or quantitative reasoning, referred to as Statway and Quantway, respectively.

From the beginning, the Carnegie Math Pathways NIC has been organized around the aim of increasing the percent of students—from 5 to 50—who achieve college math credit within one year of continuous enrollment as compared to other developmental math offerings. Multiple studies demonstrate how this aim has been met and exceeded by colleges participating in the Carnegie Math Pathways (e.g., Van Campen, Sowers, & Strother, 2013; Yamada, 2017; Yamada, Bohannon, & Grunow, 2016; Yamada & Bryk, 2016). As the hub of the NIC, Carnegie worked with various researchers and practitioners to identify key drivers that were seen as necessary for achieving their aim (see Bryk et al., 2015, p. 75). Among these drivers, Carnegie singled out various "noncognitive" factors that have been shown to affect student success (see Duckworth & Yeager, 2015). These factors can include but are not limited to students' beliefs about their ability to learn math, their sense of belonging in school, their perceptions of value for learning math, and the ways in which they set goals, monitor progress toward goals, and reflect on what worked and what did not (i.e., self-regulation skills [Zimmerman, 2002]). These factors coalesced around the idea of improving students' academic tenacity and use of effective learning strategies, what would be referred to throughout the NIC as "productive persistence."

A key instructional resource in both pathways was the use of online learning systems. From the partnership's earliest conversations, Carnegie wanted to explore how the online learning systems were being used and the degree to which data from these systems could be used to measure and support students' productive persistence in Pathways classrooms. Thus, the partnership with Carnegie was chartered as an opportunity to leverage data that was collected by the online learning systems, whereby Carnegie led the coordination with faculty teaching at 2- and 4-year colleges to interpret data products and co-develop change ideas.

### *Getting Up to Speed*

In the summer of 2014, we began working with data from the 2013–14 academic year. The partnership made the early decision to focus solely on analyzing data from Statway based on the ease of extracting data from the online system, which was built on the Online Learning Initiative (OLI) platform. The online system captured each page viewed, when it was viewed, each practice item that was attempted, when it was attempted, and whether an item was answered correctly or not. Along with page views and practice items, the online system also captured time- and item-level data from assessments, referred to as “Checkpoints.” Checkpoints come at the end of “topics” and “modules,” which organized Statway content into meaningful chunks.

For this year of Statway, there were approximately 1,600 students enrolled in over thirty 2- and 4-year colleges. Across reading, practice, and assessment activities, these students generated more than 7,300,000 rows of data. Wrangling and exploring these data involved working closely with the technology team at Carnegie as well as multiple researchers who themselves had spent time wrangling and exploring multiple datasets prior to the start of the partnership. The partnership was able to jump into analyses quickly because of the prior work that had been done by Carnegie researchers. Similar to our later work with Summit, it became important to find ways of adding value as opposed to duplicating capability. As we moved into the fall of 2014, we identified several data wrangling opportunities that could open up new levels of analysis related to the online learning system data, such as units of time (e.g., sessions and days), learning activities, and curricular organizers (e.g., topics and units) that could be used to aggregate the events that students logged, whereby these different levels of analysis could open up new opportunities to measure students’ productive persistence behaviors.

In the fall of 2014, we started to explore variation in how the online system was used across individual Pathways courses. These analyses were intended to quantify the ways in which individual instructors were using the online system at the scale of the entire network. If valid and reliable metrics could be collected on how instructors were using the online system, these measures could be used to help Carnegie staff coach faculty around best practices, understand differences in course outcomes, and provide a measure of the course context on which to better understand students’ productive persistence behaviors. For this initial analysis, we examined when students completed end-of-module Checkpoints. Using the dates that students within a course completed an end-of-module Checkpoint, we created multiple visualizations that represented both within-course variation (i.e., how students within the same course are different from one another) as well as the between-course variation (i.e., how courses are different from one another). The partnership identified that courses where most students followed the intended order of modules had higher proportions of students earning a C or higher than courses where students did not use the online system at all or where students completed end-of-module Checkpoints following a variety of different orders. Figure 6.1 demonstrates one way we visualized the dates on which students completed end-of-module Checkpoints, denoted “CP” in the figure. Each Statway course section is a row, and each time a student submitted a Checkpoint for a given module is represented by a shape. The within- and between-course variation captured

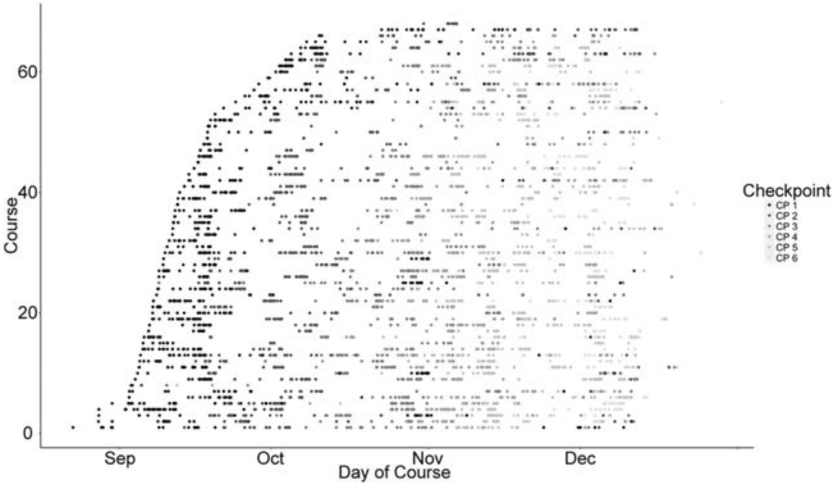


*Figure 6.1* Statway End-of-Module Completion

in Figure 6.1 provided a rationale for Carnegie researchers to reach out to faculty throughout the NIC to better understand how and why they were using the system in the ways they were.

We followed up on the course level of analyses by exploring students' use of the online system by focusing on the "session" as the level of analysis. A session was defined by the online environment as the time between logging into the system and logging out, or being timed out, of the system. Just as we looked at within- and between-course patterns for these analyses, we explored within- and between-session patterns for students. In one analysis, we coded each session that a student engaged in as a string of events, which was similar to our approach with Summit's playlist events. For example, we coded page views, practice activities, and Checkpoints as V, P, and C, respectively. For example, a student could log the following strings for two separate sessions: "VVPVVPPC" and "CC." In the first example, the student began the session with a page view, engaged in both page views and practice activities during the middle of the session, and ended the session with a Checkpoint. The second example illustrates a student logging two Checkpoint events in a row. This approach helped in seeing the different ways in which students used the system and in generating features that were used in various unsupervised and supervised learning models. Features included distinctive types of sessions, such as assessment-only sessions (all "Cs"), as well as more robust sessions where students logged Vs, Ps, and Cs within the same session. Along with these different types of sessions, we observed different within-session behaviors, such as the number of activities a student engaged in prior to taking a Checkpoint. Krumm et al. (2016) describe how we later quantified these various within- and between-session features and modeled them in relation to test and grade outcomes. These analyses, as we observed in Chapter 3, were inferential in nature, which helped in providing evidence for the importance of potentially intervening on these behaviors over time. Armed with an understanding of how different Statway courses varied in their use of the online learning system and potential behavioral measures of productive persistence, the partnership began a series of design workshops with faculty participating in the Carnegie Math Pathways NIC.

### *Design Workshops*

Design workshops were geared toward providing faculty with an opportunity to share their knowledge and expertise in interpreting data products, shaping subsequent analyses, and co-developing interventions that they could later implement. In the fall of 2015, we held our first design workshop with faculty mentors, who are a group of faculty who provide support and training to instructors throughout the NIC. In working with faculty mentors, we maintained our emphasis on experimenting with how best to organize meetings between researchers and practitioners around data. The initial workshop was organized around faculty mentors generating prototype data visualizations that they could use as part of their day-to-day teaching. For the partnership, the goal was to take faculty mentors' prototype ideas and translate them into visualizations that would later be deployed in the NIC's LMS.

The workshop was anchored in a brief presentation on a handful of measures from students' use of the system. The brief presentation was followed up with multiple individual and group prototyping activities. In collaboration with Carnegie researchers, we collected prototypes from each activity and later examined them in terms of common themes and the specific data elements they required in order to generate the visualization. We presented data products on the importance of students regularly logging into the LMS; engaging in reading and practice activities; completing Checkpoints; and reading, practicing, and assessing within the same session. Key themes that emerged from faculty mentors' prototypes were the need for more longitudinal representations of students' activity in the system and better alignments between *what* students did in the LMS with *how well* students did on Checkpoints. After analyzing the prototypes that mentors developed, we observed that prototypes often required data elements and relationships among data elements for which there was limited evidentiary support—having evidence to support the importance of the underlying student behavior represented in a visualization was a criterion that we set for the overall design process. Coming out of the first design workshop were a series of prototyped visualizations and new topics and questions to explore.

Building off of the topics from the first workshop, we engaged in a focused set of analyses, which led to a new set of data products that we presented to faculty at a second workshop in the summer of 2016. For this second workshop, we reduced the prototyping elements and increased the number of faculty who participated. Data products at this second workshop addressed the order of students' end-of-module Checkpoint completions (i.e., similar to the fall 2014 analysis), students persisting in the face of challenge (i.e., whether students return to a Checkpoint after a low score), the importance of engaging in reading as well as practice activities, and students completing both topic and end-of-module Checkpoints. As with the first workshop, prototyping activities surfaced multiple questions and follow-up topics. After the second workshop, we took stock of the evidence that was accruing around the importance of students attempting and persisting until successful on the Checkpoints within the first module and how this evidence also resonated with faculty.

Using the evidence related to completing Checkpoints, the partnership began two parallel tasks. First, we began preparing for a third design workshop where we focused on developing change ideas as opposed to prototyping data visualizations. Second, we started working with a large 2-year college in co-designing and testing strategies for helping students complete Checkpoints. Using analyses that had been presented at the second design workshop, Carnegie researchers co-designed three change ideas with participating faculty. The goal of these change ideas was to get students to complete 100 percent of their Checkpoints, i.e., their "homework," for the first two modules. Overall, we referred to this task as the "homework improvement sprint." Participating faculty members were later randomly assigned to test one or more of the change ideas in their classrooms using a planned experimentation approach (Moen, Nolen, & Provost, 2012). The co-developed changes included a work-block session prior to the start of face-to-face class (W), email reminders to students about completing their homework (E), and setting due dates within the LMS (D). Using data from the LMS, we detected positive effects for the percent of completed homework assignments as well as the timeliness of students' completion. Figure 6.2 illustrates the timeliness with which students completed each homework assignment (e.g., [1] CP 1.1.3). These boxplots illustrate how homework completion rates were more timely and less variable over time for faculty who tried out a change idea (i.e., faculty not marked "C" for control or "OTH" for other, non-participating faculty), and subsequent analysis revealed the overall benefit of the different work-block conditions (Meyer, Krumm, & Grunow, 2017).

Importantly, testing the previously mentioned co-developed change ideas was framed within the NIC and as part of the partnership as a learning opportunity, whereby evidence for the effectiveness of the individual

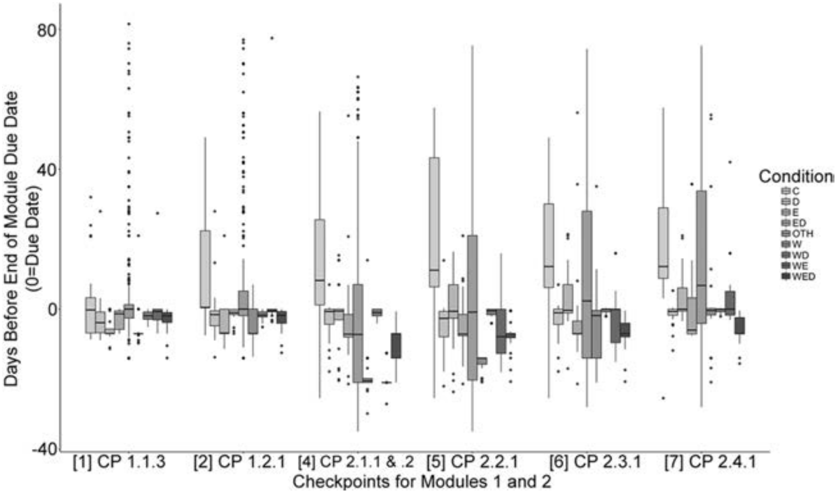


*Figure 6.2* Homework Sprint Boxplots

changes would need to be built over time as the change ideas were tried by more faculty and replicated under a variety of conditions. The ability to try out and learn from testing a change idea in multiple, diverse contexts is a core element of NICs in general and the Carnegie Math Pathways NIC in particular. Based on multiple measures generated from LMS data (e.g., the timing of submissions in Figure 6.2), we developed a robust understanding of the many aspects of the changes faculty tested out, which helped in increasing confidence that the change ideas were promising despite the comparatively few faculty who participated.

## Supporting Conditions for CDI

The two cases previously cited represent ways in which we have worked to take data-intensive research techniques to the frontlines of teaching and learning. Within each partnership, we tested particular ways of bringing researchers and practitioners together. For example, in Summit, we regularly worked with Summit's leadership teams and identified strategic opportunities to work with and learn from teachers. Similarly, in working with Carnegie, we regularly interacted with researchers and staff who supported the NIC and collectively identified the highest leverage ways of working directly with faculty and faculty mentors. Based on our experiences across these two partnerships, there are three factors that make these partnerships distinct from more traditional ways in which researchers work with practitioners: (1) research questions and topics were based on the needs of practitioners, (2) the primary audience for data products was the partnership, and (3) researchers and practitioners co-developed change ideas.

Across Summit and Carnegie, the questions and topics that were explored by the partnership were based almost entirely on the needs of practitioners or those supporting practitioners. In the case of Summit, these questions came directly from Summit leaders. For Carnegie, initial questions were based on Carnegie's interactions with faculty, and subsequent questions were generated at multiple design workshops. The primary audience for data products across both cases were members of the partnership who jointly interpreted data products, bringing with them their respective knowledge and expertise in identifying follow-on actions. Not only were researchers and practitioners jointly interpreting data products, they were also co-developing potential changes and testing them out in real learning environments. As we reflected on what was unique about these partnerships and in keeping with our design-based approach, we identified four *supporting conditions* that made each partnership successful.

In order for researchers and practitioners to come together to jointly develop data products and change ideas aimed at creating more effective learning environments, four conditions were in place:

1. The partnership between researchers and practitioners was based in **trust**.
2. An **explicit improvement method** organized multiple elements of the partnership's work.
3. **Learning events** provided opportunities for members of the partnership to collaborate and build knowledge.
4. **Common workflows** and accompanying tools supported data-intensive research, improvement activities, and project coordination.

In describing these four conditions, we do not intend to portray them as exhaustive, and it is our hope that other partnerships will use them as well as refine them over time. At a practical level, these conditions, at the very least, are intended to give future partnerships a head start in launching their own work.

### *Trust*

Across the partnerships with Summit and Carnegie, trust was a key component. Some may think trust has little to do with data-intensive research. However, if the goal is to improve educational outcomes, the role of trust between researchers and practitioners is hard to overstate (Penuel & Gallagher, 2017). Bryk and Schneider (2002) define trust as having one's expectations validated in the actions of another. Trust is a multi-dimensional construct based in respect, personal regard, competence, and integrity. Exchanging data, developing data products, and testing out ideas all benefit when both researchers and practitioners trust one another in both word and action.

Respect is experienced in the ways individuals talk to and about one another; respect is also experienced as feeling heard by other members of the partnership. At the start of the Summit partnership, for example, respectful interactions were initiated early on by listening to and building off of practitioners' questions. And across both Summit and Carnegie, respectful interactions also played a role during meetings where the partnership jointly interpreted data products. As researchers in these partnerships, we strove to create meetings—ultimately framed as learning events—where every interpretation was valued and could provide insight into understanding a data product. Acting with integrity, while a seemingly general phrase, manifested in both partnerships as adhering to promises and deadlines. Acting with integrity further involved adhering to specified procedures for working with data, as well as in keeping data analyses and change ideas focused on improving local learning environments.

Bryk and Schneider (2002) further acknowledge the importance of personal regard as a foundational component of trust. Concretely, one way in which we as researchers demonstrated personal regard involved going above and beyond in our roles as researchers; we regularly participated in last-minute presentations and conducted analyses that were not a part of either partnership's formal question development processes. While personal regard can be seen as another person going out of his or her way to help another, competence is about fulfilling one's role within the partnership. For practitioners, competence can entail understanding and describing the various learning environments that the partnership will work to improve as well as accessing and sharing relevant data. For researchers, competence means being able to carry out multiple data analysis tasks as well as being able to organize meetings and events where members of the partnership work to jointly make sense of data products. Key competencies across the two partnerships described previously were data wrangling and feature engineering, which helped each partnership merge datasets and surface new patterns and insights within their data.

In many ways, "collaboration" and "partnership" are empty words until both researchers and practitioners begin validating their words through action. Importantly, trust is a two-way street: Practitioners need to demonstrate their commitment to a partnership through both time and engagement; researchers need to similarly demonstrate their commitment by adjusting their time and schedules to better align with practitioners'. For example, at the start of the partnership, Summit highly valued working at a faster pace than that of typical research projects, and we as researchers demonstrated our ability to work at this pace. Lastly, as we observed throughout multiple meetings, trust can play an important role in jointly interpreting data products and co-developing change ideas. Trust can help mediate potentially unflatteringly outcomes that are brought to light through an analysis and trust can provide a sense of safety in brainstorming potential implications from an analysis.

### *Explicit Improvement Method*

As partnerships get started and begin to organize their work together, it can be useful to have a set of steps to follow and tools to use. Based on our work with Carnegie, we learned firsthand the important ways in which improvement science techniques can help in organizing partnership activities. In particular, strategies for understanding the problem that a partnership will work on as well as developing a theory for how to solve the problem are key steps in almost any improvement project. Across our work with both Summit and Carnegie, tools such as a causal systems analyses and driver diagrams (see Chapter 7) have all helped in shaping data-intensive analyses and co-design work. Following the development of change ideas, explicit improvement methods can be helpful in setting up iterative tests of change.

While there are many improvement methods to choose from, we have regularly made use of the *Model for Improvement* outlined by the Associates in Process Improvement, the Institute for Healthcare Improvement (see Langley et al., 2009), and the Carnegie Foundation for the Advancement of Teaching (Bryk et al., 2015). We have also used tools from the *clinical microsystems* approach developed at Dartmouth College (see Nelson, Batalden, & Godfrey, 2007). Based on our experiences, improvement science techniques play critical roles in shaping what happens before as well as after a data-intensive analysis. For example, in our work with both Summit and Carnegie, we used driver diagrams as a way of identifying key behaviors and outcomes to measure using data from digital learning environments (see Krumm et al., 2016). While improvement methods can help to shape data-intensive analyses, they are also useful in providing approaches for testing potential changes. Improvement routines, such as a Plan-Do-Study-Act (PDSA) cycle, help clarify hypotheses related to an intervention, measurement opportunities, and approaches for making sense of the test. In our work with Carnegie, both PDSA cycles and the planned experimentation methodology (Moen et al., 2012) were used to test change ideas (see Meyer et al., 2017). These approaches provided a common set of tools that members of the partnership could use in carrying out and learning from each test.

### *Learning Events*

A recurring finding from the literature on instructional improvement is that most complex interventions require practitioners to develop new skills and abilities (Cobb & Jackson, 2012). Achieving many of the goals that we set out for each partnership required both researchers and practitioners to develop new skills and abilities. The primary location for this learning occurred in meetings where members of each partnership jointly interpreted data products, co-developed change ideas, or explicitly learned from one another in a more formal setting. At a general level, learning events were structured activities where members of a partnership could develop new understandings by engaging in joint work.

Not every meeting, however, involved collaboratively developing change ideas. For example, with Summit, we organized a formal workshop where we provided direct support on using the statistical software R. Even during meetings where most, if not all, of the meeting was dominated by researchers presenting analyses to partners, very early on we came to view these not as simple information transfers from one group to another but as opportunities to demonstrate to partners the ways in which we approached problems and thought about data. In our work with Summit, this most clearly manifested in a practice where we as researchers would create data products based on mocked-up data to first demonstrate the intuition behind an analysis.

Across both Summit and Carnegie, a key feature of the ways in which we worked with partnership members as well as teachers and faculty, respectively, was continuously playing around with the genre of what it meant to meet and learn from one another. With Summit, the clearest case of this was during the data sprint, whereby in an intensive two-day event our goal was to shorten the time as much as possible from when data were analyzed to the development of explicit change ideas. With Carnegie, we experimented with different approaches for working with faculty and faculty mentors through design workshops. Each one of these latter events had a clear instructional goal and was organized accordingly.

### *Common Workflows*

A key component of data-intensive research involves analyzing, interpreting, and deriving implications from complex datasets. For these activities to take place, partnering organizations need to exchange data. In our partnerships with both Summit and Carnegie, we adopted a similar set of tools and routines for working with data from online learning and administrative data systems. First, data were queried from a database and uploaded to a password-protected, auditable, and role-based file transfer system. This system served as the central repository for raw data. Researchers were granted access to particular files; downloaded those files to password-protected and encrypted local computers; and engaged in data analysis using scriptable data analysis software. Eventually, both partnerships adopted the open-source language R and standardized many elements of the workflow using the same R packages. Data cleaning, wrangling, and analysis scripts were shared across researchers and practitioners, which created the opportunity for more reproducible analyses. Importantly, as both partnerships progressed, the added benefit of scripting every step in an analysis was that it created opportunities for practitioners to learn from worked examples and further develop their own data-intensive research skills (Gee, 2010).

Another key workflow across both partnerships involved sharing and storing data products. We experimented with collaborative file sharing services like Google Drive and Dropbox and learned over time the importance of having an intentional system in place for curating data products. In both partnerships, we produced hundreds of separate analyses—each with takeaways that informed a future action to varying degrees. Being able to revisit past work helped make partnership meetings more efficient. As we actively worked to develop better knowledge management approaches, we used improvement methods on ourselves and our own workflows. For example, our aim was to script 100 percent of a workflow and to be able to trace 100 percent of data products to a driving question. Opening up data transfer, sharing, and analysis to the tools and routines of improvement science helped us identify key opportunities for improvement and make data-intensive research activities more efficient and effective.

## Conclusion

In this chapter, we described two cases of CDI and outlined four conditions that we viewed as helping to sustain each partnership over time and ultimately turn raw data from digital learning environments into new insights and change ideas. At the outset, we oriented CDI within the broader traditions of data-driven decision making, educational data mining, and learning analytics. At the intersection of these multiple traditions, we saw clear gaps in that few partnerships existed around using data from digital learning environments and few provided detailed depictions of how to engage in collaborative data-intensive research regardless of the data source. We follow elements of the cases described in this chapter into our discussion for the ways in which CDI projects can be organized and executed across five phases in the next chapter.

## References

Allensworth, E., & Easton, J. (2005). *The on-track indicator as a predictor of high school graduation*. Chicago: Consortium on Chicago School Research.

Bailey, T., Jeong, D. W., & Cho, S.-W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review*, *29*(2), 255–270. doi:10.1016/j.econedurev.2009.09.002

Balfanz, R., Herzog, L., & MacIver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist*, *42*(4), 223–235.

Bambrick-Santoyo, P. (2010). *Driven by data: A practical guide to improve instruction*. San Francisco, CA: Jossey-Bass.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *The Journal of the Learning Sciences*, *13*(1), 1–14.

Bell, P. (2004). On the theoretical breadth of design-based research in education. *Educational Psychologist*, *39*(4), 243–253.

Boudett, K. P., City, E. A., & Murnane, R. J. (2013). *Data wise: Revised and expanded edition: A step-by-step guide to using assessment results to improve teaching and learning* (Revised and Expanded ed.). Cambridge, MA: Harvard Education Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences*, *2*(2), 141–178.

Bryk, A. S., Gomez, L. M., & Grunow, A. (2010). *Getting ideas into action: Building networked improvement communities in education*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching, Essay. Retrieved from www.carnegiefoundation.org/spotlight/webinar-bryk-gomez-building-networked-improvement-communities-in-education

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Press.

Bryk, A. S., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York, NY: Russell Sage.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13.

Cobb, P., & Jackson, K. (2012). Analyzing educational policies: A learning design perspective. *Journal of the Learning Sciences*, *21*(4), 487–521.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement*, *9*, 173–206.

Collins, A. (1992). Toward a design science of education. In E. Scanlon & T. O'Shea (Eds.), *New directions in educational technology* (pp. 15–22). Berlin: Springer.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters. *Educational Researcher*, *44*(4), 237–251.

Gee, J. P. (2010). *New digital media and learning as an emerging area and "worked examples" as oneway forward*. Cambridge, MA: MIT Press.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making*. (NCEE 2009–4067). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/

Kelly, A. E. (2004). Design research in education: Yes, but is it methodological? *The Journal of the Learning Sciences*, *13*(1), 113–128.

Krumm, A. E., Beattie, R., Takahashi, S., D'Angelo, C., Feng, M., & Cheng, B. (2016). Practical measurement and productive persistence: Strategies for using digital learning system data to drive improvement. *Journal of Learning Analytics*, *3*(2), 116–138.

Krumm, A. E., Waddington, R. J., Teasley, S. D., & Lonn, S. (2014). Using learning analytics to support academic advising in undergraduate engineering education. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 103–119). New York: Springer.

Lagemann, E. C. (2000). *An elusive science: The troubling history of education research*. Chicago: University of Chicago Press.

Langley, G. J., Moen, R., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance*. New York, NY: Jossey-Bass.

Larmer, J., Mergendoller, J., & Boss, S. (2015). *Setting the standard for project based learning: A proven approach to rigorous classroom instruction*. Alexandria, VA: Association for Supervision & Curriculum Development.

Little, J. W. (2012). Understanding data use practice among teachers: The contribution of micro-process studies. *American Journal of Education*, *118*(2), 143–166.

McLaughlin, M. W., & London, R. A. (2013). *From data to action: A community approach to improving youth outcomes*. Cambridge, MA: Harvard Education Press.

Meyer, A., Krumm, A. E., & Grunow, A. (2017, April). Are these changes an improvement? Using data to inform the improvement of homework practices. Paper presented at the *Annual Meeting of the American Education Research Association*. San Antonio, TX.

Moen, R. D., Nolen, T. W., & Provost, L. P. (2012). *Quality improvement through planned experimentation* (3rd ed.). New York, NY: McGraw-Hill Education.

Murphy, R., Snow, E., Mislevy, J., Gallagher, L., Krumm, A. E., & Wei, X. (2014). *Blended learning report*. Menlo Park, CA: SRI Education.

Nelson, E. C., Batalden, P. B., & Godfrey, M. M. (2007). *Quality by design: A clinical microsystems approach*. San Francisco: Jossey-Bass.

Penuel, W. R., & Gallagher, D. (2017). *Creating research-practice partnerships in education*. Cambridge, MA: Harvard Education Press.

Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (pp. 787–850). Washington, DC: AERA.

Piety, P. J. (2013). *Assessing the educational data movement*. New York, NY: Teachers College Press.

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, *16*(4), 385–407.

Russell, J., Jackson, K., Krumm, A. E., & Frank, K. (2013). Theories and research methodologies for design-based implementation research: Examples from four cases. In B. J. Fishman, W. R. Penuel, A.-R. Allen, & B. H. Cheng (Eds.), *Design based implementation research: Theories, methods, and exemplars: National* *society for the study of education yearbook* (Vol. 112, Issue 2, pp. 157–191). New York: Teachers College Press.

Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the science of education design studies. *Educational Researcher*, *32*(1), 25–28.

Turner, E. O., & Coburn, C. E. (2012). Interventions to promote data use: An introduction. *Teachers College Record*, *114*(11), 1–13.

Van Campen, J., Sowers, N., & Strother, S. (2013). *Community college pathways: 2012–2013 descriptive report*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Yamada, H. (2017). *Do effects of Quantway® persist in the following year? A multilevel propensity score approach to assessing student college mathematics achievement*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Yamada, H., Bohannon, A., & Grunow, A. (2016). *Assessing the effectiveness of Quantway®: A multilevel model with propensity score matching*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Yamada, H., & Bryk, A. S. (2016). Assessing the first two years' effectiveness of Statway®: A multilevel model with propensity score matching. *Community College Review*, *44*, 179–204.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, *41*, 64–70.

Chapter 7

# Five Phases of Collaborative Data-Intensive Improvement

In Chapter 6, we described two cases for how researchers and practitioners can come together to collaboratively analyze and take productive action using data from digital learning environments and administrative data systems. When we launched the partnerships with Summit and Carnegie, we sought to develop a process that other research groups could adopt and follow. Our reasoning at the time was simple: We wanted a clear process to help guide our work, there were few examples to be found, so we set out to develop our own over time. Building off of the supporting conditions described in Chapter 6, this chapter describes the outcomes of those efforts and outlines a five-phase approach for organizing a collaborative data-intensive improvement (CDI) project. Figure 7.1 illustrates each phase and the key activities within each phase. The logic behind each is as follows. Phase I involves setting up a partnership, from identifying key members to jointly defining the aim of the partnership. Phase II entails developing an overarching theory for how the partnership will reach its aim. Phase III is where the data-intensive research workflow introduced earlier fits within a CDI project—the aims and theory from Phases I and II shape data wrangling, exploration, and modeling. Phase IV is where insights from data-intensive analyses get translated into change ideas through iterative, collaborative design. Lastly, Phase V is where members of a partnership test out change ideas in real learning environments and improve upon the change ideas over time. In the remaining parts of this chapter, we describe each of these phases in further detail and follow our hypothetical high school introduced in Chapter 3 across steps and phases.

## Phase I

Organizing a partnership for success involves identifying project team members, clarifying problems the partnership is trying to solve, and specifying aims for the partnership. Research–practice partnerships are best formed around pressing needs and challenges experienced by practitioners (Coburn, Penuel, & Geil, 2013). Oftentimes, partnerships begin

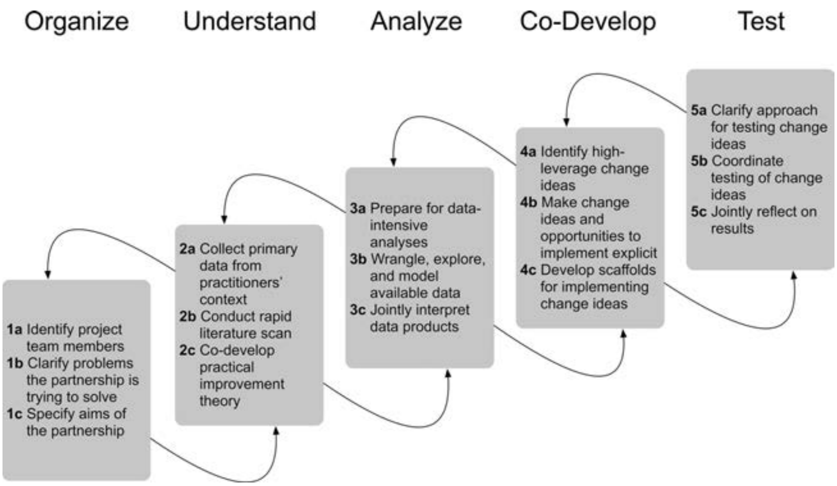

Organize
Understand
Analyze
Co-Develop
Test
1a Identify project team members
1b Clarify problems the partnership is trying to solve
1c Specify aims of the partnership
2a Collect primary data from practitioners' context
2b Conduct rapid literature scan
2c Co-develop practical improvement theory
3a Prepare for data-intensive analyses
3b Wrangle, explore, and model available data
3c Jointly interpret data products
4a Identify high-leverage change ideas
4b Make change ideas and opportunities to implement explicit
4c Develop scaffolds for implementing change ideas
5a Clarify approach for testing change ideas
5b Coordinate testing of change ideas
5c Jointly reflect on results

*Figure 7.1* Five Phases of a CDI Project

around issues that engender emotional responses due to unsatisfactory conditions or outcomes (Gomez, 2016). Going from a general sense of a need to a well-specified aim that the partnership will collectively work toward is the purpose of Phase I. Unlike a broad issue that animates people to act but does not set a direction, an *aim* is a quantifiable focus for improvement that sets the overall direction for a project (Langley et al., 2009). Before setting an aim, however, a partnership often needs to develop an understanding of the problem it is going to address. Focusing on problems at the outset of a partnership and collectively working to best understand the conditions leading to the problem can help in avoiding the familiar tendency of creating solutions first and searching for problems second (Bryk, Gomez, Grunow, & LeMahieu, 2015). In the following sections, we describe three key steps involved in Phase I.

### *Identify Project Team Members*

For a CDI project, there are often multiple roles that need to be filled and key organizational members that need to be coordinated with: champions, practitioners, data stewards, researchers, organizational leaders, and stakeholders. A key lesson from research on instructional improvement more generally is that it often takes the collective action of multiple individuals working in a coordinated way to effect change in learning environments (e.g., Cohen, Peurach, Glazer, Gates, & Goldin, 2014). A key role to identity within each partnering organization is a *champion*, who is the primary point of contact for big decisions, helps to ensure that tasks are completed on time, and keeps members of the partnership focused on key activities. Importantly, champions encourage members of the partnership to regularly update their assumptions about the project's direction and the degree to which trust is being built and sustained (Spurlock & Teske, 2015). In addition to champions, a partnership depends upon a core set of individuals who regularly attend meetings across each phase of the work and engage in specific project activities. Continuity is key, and we have observed that partnerships, especially at the outset, benefit from a consistent group of *practitioners* and *researchers*. Practitioners are individuals who work directly with learners (e.g., teachers) or are those who directly support other practitioners in working with learners (e.g., building-level leaders or central office staff who work directly with teachers). Figuring out which practitioners will attend regular meetings is not a trivial task; throughout a project, practitioners who can regularly attend partnership meetings can become the *de facto* voice for multiple constituencies.

In our partnership work, we typically take on the *researcher* role. As we have noted previously, a researcher brings to a partnership multiple skills in preparing for and conducting data-intensive analyses along with developing and testing change ideas that are informed by an analysis. Early on in a partnership, it can be useful to think of a researcher as someone who devotes his or her data analysis work to data wrangling and exploration. As we observed in both of our partnerships with Summit and Carnegie, there can be tremendous value in merging once-disparate data sources. Wrangling and exploring takes on further value as insights from theory and practice are brought together in Phase II of a CDI project and in feature engineering and predictive modeling in Phase III.

A *data steward* is a key role within an educational organization because of his or her access to data, such as databases for digital learning environments or administrative data systems. A data steward helps the partnership understand what data are available and provides the partnership with updates of data as the project progresses. Data stewards can also play a role in negotiating access to data that are collected and stored by digital learning environments but not directly held by the educational organization. In some partnerships, the data steward might regularly attend project meetings; in others, the data steward supports the project only when technical expertise is needed. Regardless of the intensity of participation, a data steward should help to ensure the secure transfer of data between organizations, such as between a school and a research organization, and can be a valuable resource in understanding the history of data systems as well as data quality issues that are present in almost every system.

Along with identifying champions and core team members, we have found that it is important to identify and coordinate with *organizational* *leaders* at both the site (e.g., principals) and district levels (e.g., central office staff or assistant superintendents). Organizational leaders can be crucial gatekeepers to resources and can play a supporting role in implementing as well as scaling change ideas. In addition, *stakeholders* can be important team members as they are most affected by the actions of the partnership. Stakeholders in educational organizations can include students and parents, and depending upon the purpose of the project, they can be regular or intermittent partners. One way in which we have interacted with students in a CDI project is through interviews and focus groups. Research–practice partnership models described by McLaughlin and O'Brien-Strain (2008), for example, outline additional ways in which parents and community members can be brought into a partnership to provide their perspectives on learners, problems to be solved, and opportunities to improve.

### *Clarify the Problems the Partnership Is Trying to Solve*

As multiple improvement science researchers and practitioners note, projects should be based on a partnership's understanding of the problem facing practitioners (Nelson, Batalden, & Godfrey, 2007). Clarifying the problems that the partnership is working to solve entails moving from the issues that initially brought potential partners together to discussions on specific processes, norms, and structures that may be contributing to the problem. There are several activities that we have drawn on in helping us to clarify problems with our partners, and one of the most useful has been a *fishbone diagram*, which is also called a root cause analysis, causal systems analysis, or Ishikawa diagram. A fishbone diagram places a problem to be addressed at the "head" of the diagram (see Figure 7.2). Above and below the line originating from the head, broad categories of factors contributing to the problem are placed. Example factors can include people, processes, materials, and norms; they are intended to be general, with increasingly more specific elements of each factor placed underneath. Connecting categories to the "spine" are perpendicular "ribs" that provide a space for capturing more specific causes. Ribs can continue to branch out in a perpendicular fashion as deeper factors are identified.

Figure 7.2 represents a fishbone diagram for our hypothetical high school introduced in Chapter 3. Recall that the problem practitioners were working to better understand were the large number of students who earned a C– or lower in the first year of the course. Practitioners brainstormed four initial categories: (1) students' study habits and strategies, (2) course design, (3) access to resources, and (4) schedule. Within these broad categories, members of the partnership identified the negative effects of students starting off track or falling behind. One potential reason for falling behind was that students did not know how to effectively

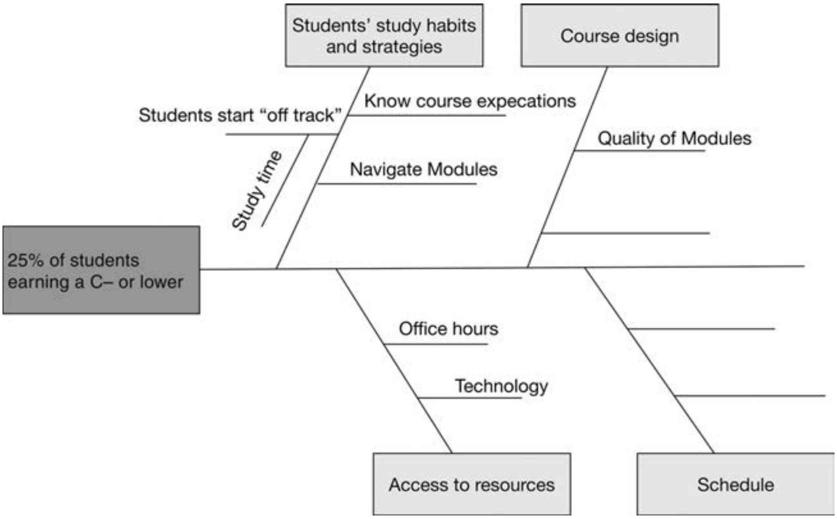


*Figure 7.2* Fishbone Diagram

use their study time in and out of school. Multiple topics such as knowing course expectations for a dual-enrollment course, navigating the modules within the learning management system (LMS), and the overall quality of the modules themselves surfaced as broad categories of potential causes for high numbers of students earning a C– or lower in the course. Mapping potential causes helped the partnership clarify which factors they wanted to prioritize in their improvement work. As we demonstrated in Chapter 3, this problem clarification work can help in providing specific questions to explore using data-intensive research techniques.

While there are no guarantees that engaging in causal systems analysis and completing a fishbone diagram will surface the right problem and the right causes, the simple act of having partners think deeply and explore multiple facets of a problem can help set a partnership on the right track. Another added benefit of engaging in a causal systems analysis is captured in the word "systems." Moving away from simple solutions and appreciating the ways in which undesirable situations are a function of multiple, interrelated factors can both reveal the true complexity of a problem as well as help a partnership prioritize factors to begin working on.

### *Specify Aims of the Partnership*

Setting concrete aims is an initial set of activities where practitioners and researchers can come together to begin integrating their respective knowledge, skills, and experiences (Coburn et al., 2013). As the partnership identifies members' roles and develops a better understanding of the problems it is trying to solve, an aim statement can become a useful resource in organizing the next steps for the collaboration. A quality aim statement answers the question "What are we trying to accomplish as a partnership?" Being clear on what the partnership is trying to accomplish can help in identifying what a partnership might need to do differently in order to achieve its aim. Said differently, a clear aim can put a partnership in a position to develop a plan necessary for achieving its aim. The causal systems analysis done earlier can be useful in this task: Problems or undesirable circumstances anchor causal system analyses; aims are often positive versions of a problem with targeted benchmarks and timelines.

A quality aim statement specifies what a partnership is working toward, the degree of improvement sought, and a timeline by which the aim will be accomplished. While there are various models for aims, such as SMART goals (specific, measurable, achievable, realistic, and time bound), we have found that answers to the simple question "what, by when?" are often sufficient in specifying an aim. Importantly, an aim statement should naturally flow from the causal systems analysis done previously, and the partnership, as a whole, should work to develop consensus around the aim statement. Bryk et al. (2015) discuss the benefits of multiple versions of a similar aim: an aspirational version and a technical version. An aspirational aim addresses the broader problem that a partnership is working on, which can provide a direction to a partnership, help in motivating members of a partnership, and serve as a reminder that a partnership is working on important problems. While aspirational aims help in motivating individuals, technical aims help a partnership measure progress toward a specific future state.

For example, in our hypothetical case high school, the aspirational aim focused on "100% of students, 100% of credit." The technical version answered the "what, by when?" question over a two-year period and acknowledged the full complexity of the problem that surfaced during the causal system analysis and available resources: "At the end of the 2019–20 school year, 95% of students who enroll in the dual-enrollment math course will earn a C or higher."

## Phase II

Understanding the best way to achieve a partnership's aim involves collecting primary data from practitioners' context, scanning pre-existing research, and co-developing a practical theory that will guide improvement work. As partnerships discuss issues, problems, and goals, multiple perspectives will emerge. Often, these perspectives are informed by the knowledge and information that individuals contributed to Phase I activities. The purpose of Phase II is to expand upon these perspectives by collecting new information grounded in both practice and research. The location for bringing these multiple sources of information together is a *practical improvement theory*, described in a later section, which can help in shaping subsequent data-intensive analyses as well as co-development work around potential change ideas for practitioners to test in their classrooms. Next, we describe three key steps involved in Phase II.

### *Collect Primary Data From Practitioners' Context*

Collecting data from practitioners' context provides an opportunity for a partnership to learn about what is happening in real learning environments. Going into Phase II, a partnership has identified a key problem of practice, made conjectures regarding the factors contributing to the problem, and clarified an aim worth pursuing. Collecting data from practitioners' context can help to confirm or disconfirm factors thought to contribute to a problem. Furthermore, collecting data from practitioners can help in developing an understanding of the key processes that make up teaching and learning. Collecting data from learning environments is particularly important for making sense of data from digital learning environments. It is easy to draw the wrong conclusions from data without an understanding of the broader instructional activities that may have contributed to what students were doing and why they were doing it (e.g., Murphy et al., 2014).

In many ways, collecting data from practitioners' context is about cataloguing and understanding how particular processes play out in an environment. A process is "a series of related work activities that together transform *inputs* into *outputs* for the benefit of someone" (Nelson et al., 2007, p. 299). Within complex systems, such as schools, there are multiple processes that build on one another in intricate and often opaque ways. To help make processes less opaque, partners can engage in *process mapping* as a structured way of articulating key steps and decisions that make up the process. Process mapping is "a method for creating a diagram that uses graphic symbols to show the steps and the flow of a process" (Nelson et al., 2007, p. 299). A central tenet of improvement science is that improved outcomes come only through improved processes. Making a process explicit can help in identifying where key processes break down, are wasteful, or are needlessly complex.

Process mapping typically begins by gathering information from practitioners through surveys, interviews, or observations with the goal of understanding how a process unfolds. This can also be accomplished by gathering a group of practitioners to discuss the sequence of steps and decisions involved in an activity or task. Within a partnership, process mapping can entail collaboratively constructing flowcharts by identifying steps, drawing boxes around each step, and connecting boxes using arrows. If a step in a process represents a decision, this decision point is signified with a diamond shape. The beginning and end of a map is signified with an oval. Using these basic building blocks, a process map highlights how well a group actually understands a process—steps that are not understood well are an invitation to learn more.

The simplified process map represented in Figure 7.3 demonstrates how teachers at the case high school worked to understand the ways in which they made students aware of "course expectations," which was an idea that surfaced during the development of the fishbone diagram (see Figure 7.2). The thinking behind the issue involved the degree to which all students knew about the required summative assessments and the importance of the assessments toward their grades. In mapping how teachers introduced course expectations on the first day and launched work on Module 1, they recognized that they discuss the grade policy and the syllabus but that they don't introduce the LMS and where the summative assessments are located in the LMS prior to starting work on course material.

### *Conduct Rapid Literature Scan*

A rapid literature scan is a focused way of helping a partnership learn about the problem it is trying to solve based on the findings and experiences of other researchers. Moreover, a rapid literature scan can help clarify what data to attend to and analyze as well as potential change ideas to test in Phase V. As Park and Takahashi (2013) outline, rapid literature scans can be pragmatic dives into the pre-existing literature to advance

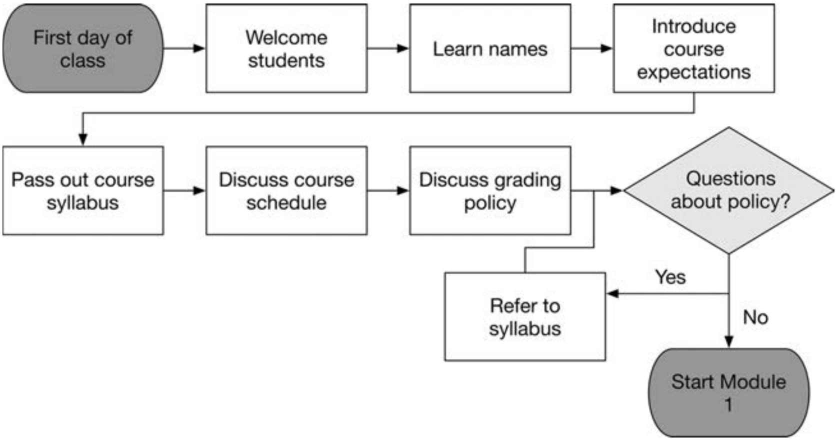


*Figure 7.3* Process Map

the learning of a partnership on a specific topic. A specific purpose or topic anchors a literature scan. After identifying a topic, a scan should be bounded by a timeline, such as 90 days, to help discipline the overall process. Along with timelines, it can be important to have a specific objective for what will be delivered at the end of the scan (e.g., a framework, annotated bibliography, sample measures, change ideas). An expedient way to begin a scan is by interviewing recognized experts; Park and Takahashi (2013) recommend that rapid literature scans strike a balance between interviews and readings. Further, they recommend a set of activities for the first 30 days, for example, of a 90-day scan. The first 30 days should focus on *scanning*, which involves conducting initial interviews and article identification as well as adjusting, if necessary, the proposed deliverable. The second 30 days involve getting feedback on what was produced during the first 30 days and continuing to make progress on the proposed deliverable as more is learned. The final 30 days involve getting additional feedback and finalizing the end product for the full partnership's review and critique.

In our work with Summit, described in Chapter 6, we engaged in a rapid literature scan to better understand research on students' self-directed learning behaviors. We consulted prior research to identify candidate measures of self-directed learning (e.g., persistence and wheel-spinning) as well as potential change ideas (e.g., messages sent directly to students through the Summit Learning Platform). Similarly, in our work with the Carnegie Math Pathways, we engaged in a rapid literature review for the purpose of developing practical measures of productive persistence (Krumm et al., 2016a). We used the assessment framework known as Evidence Centered Design to identify potential constructs, meaningful tasks from which those constructs could be measured, and the potential evidence that could be gathered from the digital learning environment used as part of the Carnegie Math Pathways (Mislevy, Steinberg, & Almond, 2003; Mislevy, Behrens, DiCerbo, & Levy, 2012). Our own rapid literature review, along with prior work done by researchers at the Carnegie Foundation, helped the partnership select what data to analyze and how to initially set up feature engineering tasks.

### *Co-Develop a Practical Improvement Theory*

A practical improvement theory is a visual representation of a partnership's approach to achieving an aim. Yeager, Bryk, Muhuch, Hausman, and Morales (2013) define a practical improvement theory as an "easily interpretable conceptual framework of the system that affects student outcomes, that practitioners view as useful in guiding their work, and that remains anchored in the best available empirical research" (p. 19). One goal of a practical improvement theory is to motivate and guide improvement work. It is *practical* because it should be used to guide local action—as opposed to supporting broader and more generalizable theory building—and because it is tentative and open to revision. As will be described later, evidence collected from practice, data-intensive work, and tests of change ideas can all be used to revise an improvement theory and a partnership's overall understanding for how to achieve a partnership-defined aim.

A popular approach for graphically displaying an improvement theory is in the form of a driver diagram, which comprises an aim, primary drivers, secondary drivers, and change ideas. In general terms, a driver diagram "consists of a team's shared theory of knowledge—which is developed by consensus—and includes relevant beliefs of team members about what must change and which ideas about how to change may result in improved outcomes" (Bennett & Provost, 2015, p. 39). As a visual tool, a driver diagram illustrates key elements of a system that need to be changed in order to achieve an aim. These key elements are aligned to specific change ideas; this alignment entails an explicit hypothesis: "If we make this change, it will affect this driver (i.e., primary or secondary), and if this driver improves, we will make progress toward our aim."

An important first step in developing a driver diagram is revisiting the aim statement developed during Phase I. From this aim, team members identify three to five necessary conditions, or primary drivers. These primary drivers are informed by the causal system analysis conducted in Phase I and the primary data collection and rapid literature scans carried out earlier as part of Phase II. Secondary drivers provide more detail around what, where, and when a primary driver will be improved upon; they provide a degree of specificity for targeting a primary driver through a specific change idea. Identifying secondary drivers, therefore, involves working backward from a primary driver to specify where and when a primary driver can be acted upon.

Figure 7.4 presents a driver diagram for the high school we have been following in this chapter. The aim for the diagram is based on the more technical aim that was specified in Phase I. Using what was learned during Phases 1 and II, the following primary drivers were selected: **Students start and stay on track**, **Students interact with peers and teachers outside of class**, and **Students interact with quality resources within Modules**. These drivers where identified as necessary for achieving the aim of having 95 percent of students earn a C or higher. While primary drivers specify what a partnership believes is necessary for achieving an aim, secondary drivers identify more specific points of intervention related to a primary driver. For example, **Students are made aware of course expectations**, is a concrete point of intervention, and as a partnership further builds out its driver diagram, it can attach specific change ideas to secondary drivers. In the case of making students aware of course expectations, a specific

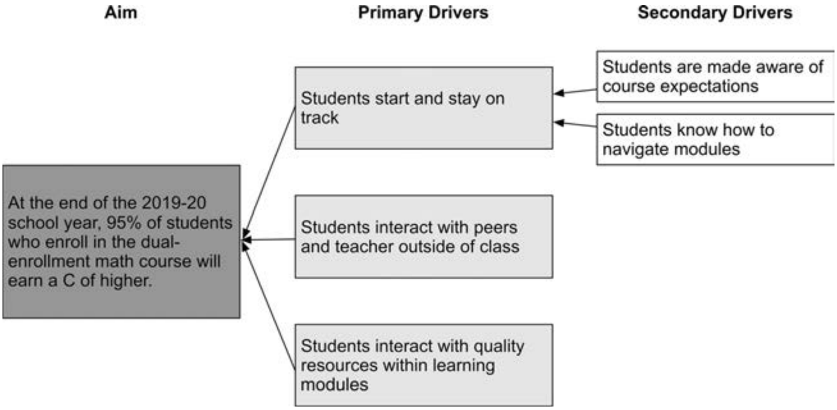

Aim
Primary Drivers
Secondary Drivers
At the end of the 2019-20 school year, 95% of students who enroll in the dual-enrollment math course will earn a C of higher.
Students start and stay on track
Students interact with peers and teacher outside of class
Students interact with quality resources within learning modules
Students are made aware of course expectations
Students know how to navigate modules

*Figure 7.4* Driver Diagram

change idea entailed developing and using a checklist to help ensure that instructors introduced specific elements of the LMS before working on course material.

## Phase III

Phase III involves preparing for data-intensive analyses; wrangling, exploring, and modeling available data; and jointly interpreting data products. Phase III of a CDI project is nothing more than the data-intensive research workflow introduced in earlier chapters. In Chapter 2, we introduced the overall workflow and how the steps worked in concert with one another. In Chapter 3, we introduced specific tools and examples for the wrangle, explore, and model steps. In this section, we describe how the workflow fits within a CDI project.

Within a CDI project, the purpose of data-intensive analyses is to develop and communicate *practical* data products that help a partnership develop a better understanding of the local education system, identify predictive relationships, and assess changes (Solberg, Mosser, & McDonald, 1997; Yeager et al., 2013). In *developing a better understanding of a local system*, practical data products can help a partnership appreciate previously unknown relationships between processes and outcomes. Based in the example of the hypothetical high school, visualizing when students first passed the summative assessment, especially as compared against the grades students earned, was useful in demonstrating previously unknown relationships. Both inferential and predictive modeling helped in *establishing predictive links* between the dates students first passed summative assessments and their eventual grades. Formal predictive modeling led to a model based in when students completed the second module (i.e., **mod_2**). Subsequent change ideas directed at improving when students complete certain modules used the day of the school year that modules were completed to *assess the effectiveness* of those change ideas. Our description of the "homework improvement sprint" conducted with the Carnegie Math Pathways from Chapter 6 offers yet another set of examples for the ways in which data-intensive analyses within a CDI project can be used to learn, predict, and assess changes.

### *Prepare for Data-Intensive Analyses*

Preparing for an analysis, as we described in Chapter 2, involves developing a research question as well as getting to know the data that will be used in subsequent analyses. Within a CDI project, developing a research question involves referring back to the various products that have been developed, such as fishbone diagrams, process maps, and driver diagrams. These products can provide direction on topics to explore, and in some cases, what data to attend to and how they might be analyzed. A driving research question can help in providing direction to a data analysis, which can reduce the likelihood of aimless data exploration. The second set of activities that can help in preparing for a data analysis involves getting to know a technology, how a technology is used, and the ways a technology collects and stores data. Getting to know a technology involves, in the case of a digital learning environment, logging into the system and exploring the various tasks and activities. It also involves seeing how, again in the case of a digital learning environment, students in classrooms are expected to use it as well as actually use it. And learning about how a technology collects and stores data entails comparing one's observations from interacting with the technology and seeing how it is used in classrooms with data dictionaries or sample database queries.

### *Wrangle, Explore, and Model*

Chapters 2 and 3 covered the topics of exploring, wrangling, and modeling data and the cases outlined in Chapter 6, in particular, highlighted the overall importance of data wrangling and exploration to a CDI project. As we illustrated in Figure 2.3 in Chapter 2, many of the steps involved in analyzing data are overlapping and the same analytical technique can serve multiple purposes (e.g., one can use inferential and predictive models to explore one's data). These three steps can consume a disproportionate amount of time and energy in a CDI project. We use the term "disproportionate" intentionally in that data-intensive analyses are a set of steps within a broader set of phases that, combined, can support positive changes in learning environments. A critical element to carrying out these steps is having a common workflow based in common tools and software—especially as partnerships move more and more into collaboratively engaging in data analyses where practitioners adopt the roles and routines of data scientists and researchers. Over time we have consolidated our tools to include the open source language R and several R packages referred to as the "tidyverse." R and related packages offer free and flexible ways of approaching each step in the data analysis process as well as tools for sharing and communicating data products, such as Markdown files and Shiny applications, both of which can be used in a meeting where researchers and practitioners come together to jointly interpret data products and brainstorm change ideas.

### *Jointly Interpret Data Products and Brainstorm Change Ideas*

After data products have been developed that address a driving question, it can be useful to convene members of the partnership to interpret data products. By "interpret" we mean noticing elements within a data product and connecting them to one's prior knowledge (Weick, 1995). Interpreting a data product involves answering the question: "What does this mean?" The different experiences and prior knowledge of researchers and practitioners can shape what they notice in a data product as well as the connections they make, and ultimately the meaning they make of it. The meaning one makes of a data product can, in turn, shape the implications one derives and ultimately the actions one takes (Coburn & Turner, 2011).

In setting up a meeting where researchers and practitioners jointly interpret data products, it is beneficial to revisit the various individuals identified in Phase I: champions, practitioners, data stewards, researchers, organizational leaders, and stakeholders. Identifying who should join in the process of interpreting data products is consequential, as individuals bring with them different knowledge, skills, and abilities, and the mix of participants can influence the interpretations and implications that are developed.

Organizing a data interpretation meeting requires different degrees of preparation depending on who is likely to attend and the overall purpose for the meeting. For meetings that involve more complex data products, we have found it beneficial to put together *one-pagers* that describe the purpose of the analysis, a sample data product, and rules of thumb for interpreting the data product. These one-pagers are useful in preparing for and focusing a meeting. While a one-pager can help prepare participants for the meeting activities, we also work to frame the data-intensive analyses in two ways during a meeting, as we described in Chapter 6. First, we describe the history of an analysis: the questions we are answering, how we got to these questions, and why the partnership might find an analysis valuable. Second, we create instructional data products that take a complex data product, distill it down into its simplest units, and present it—often using made-up data that is easier to interpret than products using actual data.

It is important to have a plan, or routine, in place for jointly interpreting data products that are presented to the group. For example, we regularly use the "I Notice/I Wonder" routine from DataWise (https://datawise.gse.harvard.edu/). We typically draw on this routine while we are walking through a data product that uses a partner's data. "Wonderings," for example, can serve as useful fodder for developing implications and potential change ideas. In brainstorming potential changes, it can be useful to have specific strategies in place, such as specific brainstorming approaches, to gather and organize potential ideas. In Phase IV, these potential change ideas can be revisited and some might be turned into explicit tools and routines.

## Phase IV: Co-Develop Change Ideas

With potential change ideas coming out of Phase III, the next phase in a CDI project is about further developing these ideas, selecting those to later implement, and making sure the necessary supports for implementation are put into place. In some cases, selecting, developing, and implementing a change idea can be easy because the change idea is relatively simple and inexpensive, such as having Statway faculty set due dates for online homework assignments as described in Chapter 6. Other change ideas, such as providing tailored feedback messages within a digital learning environment, can be more costly to develop and implement.

A key element in moving from identifying to developing a change idea is continuing to engage both researchers and practitioners in the process (Penuel, Roschelle, & Shechtman, 2007). Important aspects of fleshing out a change idea involve *elaboration* and *scaffolding* (Cohen & Ball, 2007). Elaboration is a strategy for making a change idea explicit:

> From one angle, extensive elaboration seems essential to illuminate an innovation's requirements for use, to alert designers and implementers to work to be done, and to reveal potential problems. Less-elaborated designs would be not only less useful but even self-defeating, for they tacitly delegate large amounts of invention to implementers, increasing the probability that the implementers would interpret interventions as versions of conventional practice, since the designs offer little guidance for anything else, and conventional practice is both familiar and understood by implementers.
>
> (p. 25)

Scaffolding relates to the supports that are put into place to help practitioners implement the change idea. Scaffolds can include everything from formal professional development sessions to worked examples that help a practitioner see what the change idea looks like in practice. Co-developing change ideas, therefore, can involve multiple steps, such as selecting high-leverage change ideas from the multiple brainstormed ideas, making the selected ideas explicit, and developing scaffolds to help practitioners learn about the change idea and implement it.

### *Identify High-Leverage Change Ideas*

High-leverage change ideas are those where there is evidence to suggest that a small effort may lead to large improvements. The partnership's improvement theory is one tool for beginning to identify high-leverage change ideas. The primary drivers outlined in the theory are themselves intended to represent high-leverage aspects of the system that the partnership can modify to achieve its aim. Thus, the practical improvement theory developed during Phase II not only helps in shaping the data-intensive analyses, but also helps in prioritizing potential change ideas.

Along with a partnership's practical improvement theory, more general criteria for selecting change ideas include the following: short lead time, low cost, and control (Nelson et al., 2007, p. 326). Partnerships can be buoyed by early wins. Selecting a change that can be implemented with a short lead time increases the odds of attaining an early win by selecting something that will be easy and fast to implement. Change ideas with a short lead time are also typically low cost (i.e., in terms of time and money). Having control means that the partnership does not need extensive permissions to try out the change idea. Starting with a change idea that has all of these features means that the collaborators can become accustomed to moving through improvement processes first, before tackling more difficult change ideas.

In addition, as Yeager et al. (2013) observe, evidence from research can be useful in selecting change ideas, particularly those studies identified during the rapid literature scan of Phase II. As partnerships use their practical improvement theories and general criteria in selecting changes, Nelson et al. (2007) note the importance of having those who will be implementing a change participate in the selection of the change to implement as well as the process of fleshing out the change idea. Based on our own experience, meaningful change ideas coming out of a data-intensive analysis sometimes involve simply collecting more data in order to understand a problem better. For example, in our work with both Summit and Carnegie, we engaged in multiple data product development and interpretation cycles and identified the need to identify as well as gather more evidence before launching into a change effort for the practitioners to implement.

### *Make Change Ideas Explicit*

Understanding and implementing explicit processes are critical for achieving reliable performance (Bryk et al., 2015; Nelson et al., 2007). Explicitness makes what is expected of a practitioner concrete (Krumm et al., 2016c) by detailing the steps and decisions involved in a process, which is similar to the idea of a process map. However, instead of just naming a step, as in a process map, an explicit change idea includes a description of what each step entails. For example, in a recent project where we worked with a group of teachers to increase the quality of science discourse in elementary classrooms as our improvement aim, we broke down the steps involved in conducting a whole-class discussion and added content-specific phrases that a teacher could use that were unique to the lesson. Instead of telling teachers to "lead better discussions," we worked with them to develop explicit, elaborated protocols on how to actually go about leading a discussion in a way that drew on the expertise of both researchers and practitioners. Another key component of explicitness, along with having steps and decisions outlined, is clarifying the situation or context in which a practitioner should implement the change idea—for example, tailoring the protocol to specific science lessons and clarifying when during a lesson to use the protocol (Moorthy & Krumm, 2017).

There are multiple ways to make change ideas explicit. In leading the homework improvement sprint described in the previous chapter, an improvement coach from Carnegie worked with faculty to select potential change ideas and to make them explicit for testing. Selecting changes to develop and implement proved easier than making the changes explicit. For example, the idea of sending email reminders to students seemed simple and straightforward. However, the timing as well as the substance of emails proved more difficult to agree upon. Efforts to create explicit tools and routines surfaced deeper beliefs related to the purpose of homework and the responsibility of students in managing their own workloads. Thus, in the process of making change ideas explicit, a partnership can wrestle with big issues and uncover additional aspects of a problem the partnership is working to solve (Meyer, Krumm, & Grunow, 2017).

Another strategy for making a change idea explicit is to iterate on it by having one group within the partnership brainstorm how the change could be implemented and then turn their ideas into a prototype. The prototype can then be handed over to another group for their additions. Over time, an implementable change idea can emerge from this process. In our work with Carnegie, this approach helped in developing more resource-intensive change ideas, such as new visualizations of productive persistence measures that could be implemented in the online learning environment. With each iteration of the prototype, the partnership got more specific about the visualizations and the resources needed to develop and test them.

### *Develop Scaffolds to Support Implementing Change Ideas*

While important, explicitness is not a substitute for working with practitioners on what they need to know in order to implement a change idea well. Some changes may not require a lot of up-front learning on the part of practitioners. However, we have observed that even simple change ideas can require, for example, reminders to implement them. Professional development and reminders are both examples of potential scaffolds that partnerships often need to put into place in order to implement a change idea successfully. Thus, just as the change ideas themselves need to be made explicit, so too do the different supports and scaffolds that the partnership will provide to practitioners in order to implement the change idea.

## Phase V: Test

Getting to the point where practitioners test a co-developed change idea is a tangible milestone for any CDI project, and many successful projects strive to test change ideas as quickly as possible. Some of the best learning for a partnership can occur from testing change ideas in real classrooms, and there are specific activities that can help ensure successful testing of a change idea. Success is defined by what the partnership is able to learn from a test in relation to their jointly developed aim statement and improvement theory. We use the term *improvement cycle* for efforts made by researchers and practitioners to test a change idea, which can create both short- and long-term learning opportunities for a partnership. Short-term learning involves getting the most out of an individual test by starting with an explicit idea, a hypothesis about what will happen as a result of the change, data that will help in understanding what happened, and a mechanism for collecting data used to test a hypothesis. Improvement tools, such as Plan-Do-Study-Act (PDSA) cycles, can be used to structure these tests. Long-term learning involves documenting and keeping track of the multiple tests of change over time, and importantly, synthesizing what is learned from across multiple tests.

Testing can be done at different scales and with different numbers of practitioners. Bryk et al. (2015), for example, identify the role of *confidence* and *capability* in deciding on the scale at which testing can occur. Confidence stems from the strength of the evidence base, both from practice and research, for the potential positive benefits of a change idea. Capability concerns the relationship between current and necessary knowledge, skills, and abilities for enacting a change idea. Only when confidence and capability are both high should a partnership think about trying out a change idea at a scale larger than an initial handful of willing participants.

### *Familiarize Partnership With Approach for Testing Change Ideas*

Prior to testing out a change idea, partners must agree on a specific approach and timeline for testing it. For our work with Carnegie on increasing homework completion, for example, Carnegie researchers tested the three separate change ideas by randomly assigning combinations of changes to faculty members at the beginning of a semester. This provided the opportunity to test and refine the change ideas over two cycles at the beginning of the fall and winter semesters. In other partnerships, we have worked with groups of teachers to implement and iteratively refine change ideas, using PDSA cycles on a weekly basis (e.g., Krumm et al., 2016b; Moorthy et al., 2016). No matter how the change idea is tested, those who are doing the testing can benefit from being familiar with the story of how the change idea was developed—causal system analyses, driver diagrams, and the data products produced in Phase III are key resources. Knowing the story can help those doing the testing understand the rationale for the changes being attempted, which can help in building will and in making sense of results from initial tests. Along with describing how the change idea emerged, the partnership should familiarize testers with the explicit details of the change ideas and the scaffolds that can be used to support their learning and testing.

In other scenarios, more may be asked of practitioners—from filling out formal PDSA forms to collecting data for subsequent analysis. In some situations, we have had teachers complete customized PDSA forms to document their tests of change ideas. One challenge with formality is that PDSA forms can come to be seen as added paperwork and lose their value for documenting tests. Finding ways to obtain measures of change idea implementation without burdening practitioners is a major challenge in Phase V. In the Statway test of homework completion routines, Carnegie researchers directly observed how change ideas were implemented and the partnership analyzed system log data to produce process and outcome measures. Given the variety of approaches and strategies that are available, it is important that teams are clear on an initial approach and are open to refining the approach depending upon what is helping the partnership.

### *Coordinate Tests of Change Ideas*

Coordinating tests of change ideas involves clarifying how implementers will be brought together to share what the partnership is learning. For example, regular face-to-face or virtual meetings can provide opportunities for practitioners testing change ideas to report on what they are finding and the adaptations they are making. These meetings can also harness the knowledge of a broader set of colleagues in brainstorming further adaptations. In addition to regular meetings, tests can be coordinated by an improvement coach, common forms and reporting documents, and a collaborative infrastructure for storing and sharing what is learned from a test. An improvement coach is an individual with a background in improvement science who can help practitioners plan and document their tests. Common forms and reporting documents help each practitioner collect information from a test that can later be aggregated across tests. As we noted previously, the potential downside of common forms and documents is that they can be perceived as paperwork and not completed as intended. For this reason, an important tip when using common forms and documents is to be open to modifying them over time based on practitioners' feedback on a form's relevance and ease of use (Krumm et al., 2016c). Lastly, common forms and documents, schedules, and other resources can all be stored and shared using online collaborative tools, such as an intranet, wiki, or cloud-based file hosting service. These tools can facilitate easier communication among those doing and supporting the work of testing change ideas.

### *Jointly Reflect on Results From Multiple Tests*

After engaging in multiple tests of change, it is often beneficial for a partnership to stop and reflect on what has been learned. The Institute for Healthcare Improvement's *Breakthrough Series Collaborative* (2003) describes the importance of a summative meeting following tests of change ideas and learning sessions. These summative sessions can give researchers and practitioners the opportunity to present their overall findings for a set of tested change ideas, to celebrate successes, and to plan for the next improvement project. Sharing findings across multiple tests can help a partnership more fully assess the overall project, and the partnership can strategize on longer-term co-development tasks that could not be accomplished originally during Phase IV as well as rethink potential data-intensive analyses.

Reflecting on the results from multiple tests often raises new issues that can serve as the starting points for additional improvement work. Aims can be revisited and refined, improvement theories can be added to, new data products can be developed, and new change ideas can be made explicit and prepared for future testing. Across multiple projects, we have found that joint reflection leads naturally to *continuous* improvement. After testing a series of change ideas, we frequently return to Phase III to explore the data collected during a set of tests to better understand the impacts of changes and to surface new questions for the partnership to explore.

## Conclusion

In this chapter, we described the five phases of a CDI project. These phases were identified from our work across multiple partnerships, and when paired with the conditions outlined in Chapter 6, they capture what it means to engage in CDI. The phases of *prepare* (Phase I) and *understand* (Phase II) specify the importance of setting clear aims for a partnership and using prior research and wisdom from practitioners to shape data-intensive analyses that occur during the *analyze* phase (Phase III). Joint data product development and interpretation set the partnership up for *co-developing* (Phase IV) change ideas and *testing* (Phase V) them in local learning environments. In bringing together improvement science and data-intensive research, all in the context of collaborating with practitioners, we have sought to make explicit how the trends discussed in Chapters 5 and 6 can come together in focused partnership work. In the next and final chapter, we reflect on the future of CDI.

## References

Bennett, B., & Provost, L. (2015, July). What's your theory? Driver diagram serves as tool for building and testing theories for improvement. *Quality Progress*, 36–43.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Press.

Coburn, C. E., Penuel, W. R., & Geil, K. E. (2013). *Research-practice partnerships*. New York, NY: William T. Grant Foundation.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement*, *9*, 173–206.

Cohen, D. K., & Ball, D. L. (2007). Educational innovation and the problem of scale. In B. Schneider & S. McDonald (Eds.), *Scale-up in education: Ideas in principle* (Vol. 1) (pp. 19–36). Lanham, MD: Rowman & Littlefield.

Cohen, D. K., Peurach, D. J., Glazer, J. L., Gates, K. E., & Goldin, S. (2014). *Improvement by design: The promise of better schools*. Chicago: University of Chicago Press.

Gomez, L. (2016, December). *Identifying and refining high leverage problems*. Presentation for INCLUDES Center Webinar Series. Menlo Park, CA.

Institute for Healthcare Improvement. (2003). *The Breakthrough Series: IHI's collaborative model for achieving breakthrough improvement*. IHI Innovation Series White Paper. Boston: Institute for Healthcare Improvement.

Krumm, A. E., Beattie, R., Takahashi, S., D'Angelo, C., Feng, M., & Cheng, B. (2016a). Practical measurement and productive persistence: Strategies for using digital learning system data to drive improvement. *Journal of Learning Analytics, 3*(2), 116–138.

Krumm, A. E., Zheng, Y., Biesenger, K., Moorthy, S. M., Boyce, J., Gilligan, E., Alozie, N., Miller, D., & Welch, G. P. (2016b, April). Analysis of English learners' science achievement as a boundary practice in a research-practice partnership. Paper presented at the *Annual Meeting of the American Educational Research Association*. Washington, DC.

Krumm, A. E., Boyce, J., D'Angelo, C. Podkul, T., Feng, M., Christiano, E., & Snow, E. (2016c). *Project-based learning virtual instructional coaching networked improvement community*. Final report. Menlo Park, CA: SRI Education.

Langley, G. J., Moen, R., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance*. New York, NY: Jossey-Bass.

McLaughlin, M. W., & O'Brien-Strain, M. (2008). The youth data archive: Integrating data to assess social settings in a societal sector framework. In M. Shinn & H. Yoshikawa (Eds.), *Toward positive youth development: Transforming schools and community programs* (pp. 313–332). New York: Oxford University Press.

Meyer, A., Krumm, A. E., & Grunow, A. (2017, April). Are these changes an improvement? Using data to inform the improvement of homework practices. Paper presented at the *Annual Meeting of the American Education Research Association*. San Antonio, TX.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., & Levy, R. (2012). Design and discover in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, *4*(1), 11–48.

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

Moorthy, S., Gilligan, E., Alozie, N., Krumm, A. E., Boyce, J., Welch, P. G., Miller, D., & Biesinger, K. (2016, April). Co-designing supports for science instruction: Lessons from a research-practice partnership. Paper presented at the *Annual Meeting of the American Educational Research Association*. Washington, DC.

Moorthy S. M. & Krumm, A. E. (2017, November). Improving science learning experiences in diverse elementary classrooms: Lessons from a research-practice partnership. Poster presented at *EdSurge Fusion Conference*. Burlingame, CA.

Murphy, R., Snow, E., Mislevy, J., Gallagher, L., Krumm, A. E., & Wei, X. (2014). *Blended learning report*. Menlo Park, CA: SRI Education.

Nelson, E. C., Batalden, P. B., & Godfrey, M. M. (2007). *Quality by design: A clinical microsystems approach*. San Francisco: Jossey-Bass.

Park, S., & Takahashi, S. (2013). *90-day cycle handbook*. Carnegie Foundation for the Advancement of Teaching. Retrieved from www.carnegiefoundation.org/resources/publications/90-day-cycle-handbook/

Penuel, W. R., Roschelle, J., & Shechtman, N. (2007). The WHIRL co-design process: Participant experiences. *Research and Practice in Technology Enhanced Learning*, *2*(1), 51–74.

Solberg, L., Mosser, G., & McDonald, S. (1997). The three faces of performance measurement: Improvement, accountability, and research. *The Joint Commission Journal on Quality Improvement*, *23*(3), 135–147.

Spurlock, B. W., & Teske, P. A. (Eds.). (2015). *All in: Using healthcare collaboratives to save lives and improve care*. Roseville, CA: Cyosure Health.

Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks, CA: Sage Publications.

Yeager, D., Bryk, A., Muhuch, J., Hausman, H., & Morales, L. (2013). *Practical measurement*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Chapter 8

# Lessons Learned and Prospects for the Future

The basic premise of this book is that use of data for educational improvement is nearing a tipping point enabled by the confluence of better datasets and analytic techniques and tools that make data easier to access, manipulate, and understand. But there are strong social and organizational aspects to data use—the kind of data-informed improvement work we have described in the preceding chapters will require new learning and ways of working on the part of all participants. In this final chapter, we step back from the details of data analysis and collaborative data-intensive improvement (CDI) practices and tools to reflect on the challenges that this kind of work poses for the participants who need to make it work. We consider the changes in practice that CDI requires on the parts of education researchers, data scientists, education leaders, and frontline practitioners, as well as learning technology vendors and developers. We then reflect on some of the lessons we have learned from our own and others' early CDI efforts and offer some predictions with regard to future trends. Finally, we close with an invitation to others to undertake this kind of work and contribute their own insights into how to do it most productively.

## Requirements for Changing Perspectives and Practices

The description of CDI in Chapters 6 and 7 highlights the need for additional competencies, including a deep understanding of the goals, constraints, and processes of the local education system; knowledge of the learning sciences and educational research literatures; improvement science concepts and tools; leadership skills; assessment (e.g., psychometrics) expertise; and knowledge of education research design and associated statistical approaches. This wide-ranging set of requirements reflects the complexity of the enterprise, but it should be remembered that CDI is a team endeavor. No one person needs to have or is likely to have all of these competencies. The important thing is to have them available somewhere on the team and to be able to call on them when needed.

While CDI participants do not need to possess every required competency individually, every participant is likely to need to make some changes in his or her usual ways of working.

**Education researchers** will find that CDI—like any serious research–practice collaboration—will require them to become comfortable with no longer having sole ownership of their research question, outcome measures, and research design. This is a fundamental change from the way most education researchers have been trained and from what they have come to expect. Making the shift is difficult also because many educators too are accustomed to ceding responsibility for research decisions to researchers. Researchers may find that their collaborators are reluctant to question them, which is one of the multiple reasons that trust and mutual respect are such important drivers for engaging in CDI.

Education researchers, especially those steeped in traditional research design, such as random-assignment experiments, may also be challenged by the more engineering- and design-based methods of CDI. The central goal under CDI is to improve outcomes, not to execute perfect research designs. Learning fast through small tests of change means that many of those tests will involve relatively few teachers and instructors. From a statistical view, these tests will be underpowered and hence unlikely to show statistically significant differences unless the intervention's impact is very large. Moreover, statistical significance (i.e., *p*-values) will not necessarily drive decisions about whether or not to stick with an intervention. Other kinds of information, including the insights and opinions of teachers and students, will be taken into account, and conclusions from early iterations will be tentative, awaiting more data from additional tests across expanding numbers of increasingly diverse environments.

For these reasons, among others, education researchers are likely to find that some scholarly journals will not be receptive to articles based on CDI work. Rapid improvement cycles are designed to generate evidence that is good enough to inform the next implementation cycle, which is different from adhering to the methodological requirements for publication in peer-reviewed research journals. This challenge may ease somewhat over time as conceptions of rigor and quality in education research continue to evolve. There are some indications that a more expansive view of what education research can and should be is emerging (Barab & Squire, 2004; U.S. Department of Education, 2013; Penuel & Gallagher, 2017). For example, some scholars have commented on the fallacy of defining research quality exclusively in terms of adherence to a random-assignment design supporting causal inference (e.g., Ginsburg & Smith, 2016).

It can be useful to see that CDI engagements are perfectly compatible with large experimental studies of educational impacts performed at a point in the collaboration when the improvement idea has resulted in a definable intervention that is considered ready for implementation at scale. Moreover, we would argue that experiments performed as the culmination of a CDI effort are actually more likely than others to find positive impacts by virtue of the fact that they take a systematic view of the intervention and its implementation. It is important to recognize, however, that CDI is not solely a handmaiden or onramp to large-scale experimental studies. These modes of inquiry are distinct and serve different purposes—improving learning environments over time (i.e., CDI) as opposed to building knowledge or making strong claims about an intervention's effectiveness (i.e., experimental research designs). Though, an important caveat to this claim is that the practical theory building and measurement work that accompanies a CDI project, while directed at supporting the learning of those engaged in the work, are the building blocks of a well-designed experimental study, while the inverse is often not the case. Said differently, measures developed for a CDI project can be used to better understand the results of an experimental study; data collected for the purpose of an experimental study, alone, are often not well suited for collaborative, iterative improvement work.

**Educational data scientists** participating in CDI may need to conceive of their work on longer time scales, going beyond data analysis and into intervention development and testing. Once intervention ideas have been formulated and are being tried out, the analytical work of a data scientist can be useful in providing a near-term outcome for rapid small tests of change. But data scientists might have to adjust to the expectation that they build a much deeper understanding of educational practices at their partner institutions than they may be used to. Another change for educational data scientists engaged in CDI is being called on to move beyond familiar routines for exploring a single dataset, such as from a particular digital learning environment. For educational data scientists engaged in CDI, they will need to be polyglot and quick studies with data from diverse sources used to serve diverse improvement goals. Moreover, data wrangling and exploration are often just as important as predictive modeling when it comes to adding value to practitioners.

**Education leaders and frontline practitioners** will find that CDI imposes a discipline on the way in which they make decisions around the problem of practice being worked on jointly. They are committing to a specific goal, including a metric and target level for the amount of progress toward that goal that will be considered a success. They are also committing to use data rather than instinct or political expediency to shape decisions about future instruction. Becoming a reflective and data-informed practitioner in this way requires time—both to understanding ideas coming from other fields (i.e., data science and learning science research) and to implementing the new instructional practices coming out of the collaboration. Time requirements for participating in CDI will be particularly taxing for core team members from education institutions because they typically already have full-time responsibilities.

**Technology developers** involved in this kind of research may find themselves challenged by CDI's requirements for openness. A learning technology startup company can be driven by the goal of getting a product to market as rapidly as possible and scaling to large numbers of customers in order to attract further investment. Under these circumstances, it is natural that they are concerned about the possibility of public release of any information about lack of impact that some might be construed as evidence of weaknesses in their product. Allowing data scientists, researchers, and practitioners to look under the hood and work with data from your learning system requires tremendous confidence in the quality of your product and in the good faith and discretion of your CDI partners. Investing time in building relationships with these partners is also a real cost that may worry leaner startup companies.

**All of these collaborators** will need to overcome differences in the key concepts and terminology used in their respective fields to learn to communicate effectively with one another. They also need to become aware of and sensitive to the different goals, constraints, workflows and communication styles within their respective organizations. Researchers need to appreciate the fact that practitioners have full-time jobs other than participating in the research–practice partnership. Finding time to be full-fledged partners can be challenging for them, and other partners need to respect their time constraints. An additional complicating factor is that the participation of researchers will often be funded by some organization outside of the partnership, and research funding often comes with its own requirements around timetables, products, and methods.

Examples described in Chapters 6 and 7 demonstrate that these challenges can be overcome, but doing so will be easier if the challenges are anticipated and the team actively addresses them early. And although the many differences between research organizations, learning technology companies, and education institutions can pose challenges for CDI, as described previously, there are also multiple congruencies that can support collaborations. The concept of working through multiple, iterative cycles, for example, is common to data science, technology development, and continuous improvement efforts in education. More fundamentally, all of the partners share the goal of providing students with learning experiences that will lead to more consistently positive outcomes. Shared values around this goal are an important driver for CDI, as described in Chapter 6.

## What We're Learning

Across the multiple projects and partnerships in which we have worked, we have identified the following takeaways.

### *Formulating Meaningful Variables Is Central to the Work of CDI*

The extraction of data that can stimulate insights into teaching and learning problems and their potential mitigation requires reducing and focusing a dataset. The thousands of observations that can accumulate for each student in some environments often need to be consolidated into constructs relevant to teaching and learning, which we described as *feature engineering* in Chapters 2 and 3. The learning analytics consulting firm Civitas Learning describes four categories of derived variables they have found useful in their work relating online learning activities to college outcomes (Civitas Learning, 2016). *Consistency* variables capture how regularly a student engaged in a certain course activity such as viewing learning resources, completing embedded assessments, or posting to a discussion board. *Normative* variables capture a student's performance in a course relative to that of other students in the same course. *Min and max* variables are the lowest and highest values of something, such as percentage correct on a quiz or number of log-ins in a week, for an individual student. *Average* variables for a student look across multiple datasets or time periods to compute an average for each student, such as average grade across all courses or average monthly attendance rate for a school year. Learning scientists, instructors, and data scientists all have expertise to contribute to these critical decisions about the right features to measure and how they should be defined. Defining these constructs conceptually and operationally constitutes a major task for data scientists working in collaboration with practitioners (Bienkowski, Feng, & Means, 2012; Siemens & Baker, 2012).

### *Data Scientists Need to Make Sure They're Working on a Real Problem of Practice*

For a data scientist, a large and complex dataset offers seemingly endless possibilities, and it can be difficult to resist the temptation to jump into the data prematurely, before the collaborators have agreed on what problem they're trying to solve. It's easy to find what looks like an anomaly in learning system data (e.g., a group of students who fail in an early learning module and then do extremely well in a subsequent one) and then to make conjectures about the teaching and learning problem these data might reflect. But that conjecture can easily be misguided if the partnership analyst lacks contextual information about where, when, and under what circumstances the learning system was used. Moreover, even if the conjecture is correct, the problem addressed by the partnership may well be trivial from the standpoint of practitioners. A commitment to CDI means working collaboratively between researchers and practitioners to identify the problem of practice to be solved and letting that problem drive data analyses—not the other way around.

### *You Have to Leave Your Desk to Understand Teaching and Learning*

Instruction is best understood as a complex phenomenon emerging from the interactions among students, teachers, instructional resources and a host of contextual factors (Cohen, Raudenbush, & Ball, 2003). Data from administrative systems, digital learning environments, or sensors and recording devices will not capture everything that is important about instruction and the context within which it occurs. Researchers need to supplement what they can learn from such data systems with the kind of qualitative data that can be gained from observations, interviews, and focus groups with teachers and students. Accessing a digital learning environment from afar can be useful, as are discussions with the system designers about their intentions and the system architecture, but these activities are no replacement for actually talking to students and teachers about how they use the system and how they think about its components and functionalities.

### *Data for Generating Change Ideas Should Not Be Confused With Data for Other Purposes*

CDI uses data for these three different purposes—understanding, prediction, and assessing changes—and it's important that researchers let the team's current purpose drive the analyses they run. In Phases I and II of the CDI process described in Chapter 7, data on system practices and outcomes are examined to enhance the team's understanding of the problem they want to work on and to spark ideas about how they might address it. As described in Chapter 7, we have found that at this stage a diverse set of visualizations can be useful for inspiring ideas for change. But these data products are merely a means to that end; they are not the same thing as changing what and how teachers teach and students learn. Later in a collaboration, after a partnership has generated change ideas and implemented one or more of them in small tests of change, researchers can extract data on the outcome measure that has been defined as the early indicator for their long-term goal. At this point, researchers do not need to be looking for all the variables that correlate with that outcome or to be looking for interesting student profiles. Researchers' primary job at this point is to provide the results from the outcome measure so that the team can see whether or not the change in practice enhanced the outcome the team had selected for improvement.

We can illustrate this required change in focus with a hypothetical CDI project that might have been triggered by the experiences of Georgia State University, which worked with researchers to review a decade's worth of data for students in its nursing program. Their expectation was that grades in the Conceptual Foundations of Nursing gateway course would predict program completion (Treaster, 2017). They were surprised to learn that this was not so, but grades in a nursing student's first college math course was predictive of program completion. At this juncture, a CDI team focused on improving outcomes would begin examining data on the introductory math course as conventionally taught to generate ideas about how it could be changed in ways that would result in more nursing students completing it successfully. A CDI effort around this problem of practice might then design an early intervention program for nursing students earning less than a B midway through that math course. In the next phase of their work, the CDI team would test out the intervention to ascertain the extent to which it increases the proportion of nursing students earning As or Bs in their initial math course. Here, an experimental or quasi-experimental design would be appropriate and course grade would provide the needed outcome measure. Later impact analyses could examine the extent to which the intervention also increased the long-term outcome of completing the nursing program. These data are different from the data that informed the change idea, and in this case conventional statistical approaches from education research would suffice for testing the efficacy of the intervention.

Another confusion of purpose can occur when the data products used with the core team to try to understand the problem or trigger ideas for addressing it are incorporated into the intervention. The visualizations produced to display complex datasets to try to understand a problem or predict student success were not designed to be folded into instructional practice. Visualizations commonly used by data scientists are perceived by educators as difficult to understand, and impossible to interpret with just a rapid glance. If predictive analytics to identify students who should receive a different kind of learning experience or more intensive support are to be part of an intervention, the practitioners who must act on this information need to receive it regularly and in a form that can be understood easily and quickly. Data products should be designed expressly for this latter purpose and their usability should be tested as part of the change idea.

### *You Should Set Up Data Security Procedures and Data Use Agreements Before Touching Any Individual-Level Data*

The importance of data privacy and security for CDI was discussed extensively in Chapter 4. Using large-scale data systems with records for individual students and detailed logs of student behaviors when learning online are characteristics of CDI. Because individuals' learning data are in the dataset, and data records from multiple systems may be linked, researchers have a strong obligation to shield individual identities and prevent acquisition of the data by unauthorized individuals. As described in Chapter 4, there are some well-established routines and technology tools for securing data, but applying them requires resources and vigilance. Admittedly, the first time a CDI team deals with obtaining institutional review board approval, data anonymization, secure file transfer, and so on, each of these steps is likely to feel complicated and labor intensive. But again, being forewarned is being forearmed. Anticipating these steps in project plans and data use agreements makes for a smoother workflow. We have found that having someone on the team who has specialized in executing these steps and acts as the project's data steward can make sure that human subjects and data protection functions are executed efficiently. As they become part of standard practice, the time required to execute these functions will decrease, but it will never become insignificant.

## Looking Ahead

We're under no illusion about our ability to make long-term predictions, especially in an area changing as quickly as data science. However, we do see some emerging trends that can reasonably be expected to influence CDI type work over the next five years.

### *Learning Technologies Better Designed to Support Data Analytics for Improvement*

Having worked with data from dozens of learning technology products, we've found that they vary markedly in the ease with which they can be used to inform instructional decision making. Despite the millions of data points that are potentially available, connecting an action a student takes within the learning system with the broader learning context in which the action was taken may not be possible from the log data, thus posing challenges for interpreting the data. For example, some systems log the correctness of each student's last answer to each question but not how many times the student tried to answer that question. How then do we disentangle the degree to which the content was learned from the extent to which the student persisted after giving a wrong answer? Learning software that was not intentionally engineered to support analytics yields data elements that may not be related to a theory of learning. We expect and hope that future learning technology designers will anticipate subsequent analyses and improvement efforts as they design the data collection and storage features for their software. This will require conceptualizing how content or services can contribute to learning in the form of a coherent theory for improvement (see Chapter 7), and then instrumenting the learning software to collect data that would validate, refute, or suggest directions for refining that theory.

### *Better Data Governance*

The application of data science to large datasets generated by digital learning systems started before any consensus around data ownership, openness, and appropriate data security and privacy had time to emerge. Many education institutions signed contracts with learning technology and online service providers before thinking about the value they might derive from having access to data themselves. Most of the public discussion so far has been about whether or not technology providers should be able to use system data for other purposes (e.g., to inform targeted advertising or improve their products) as discussed in Chapter 4. But some higher education institutions in particular are starting to negotiate with vendors over who has access to, and ownership of, the learning data of their students. We believe that this issue, as well as those of data privacy and security, will become a standard part of negotiations among school systems, vendors, and external research organizations. One thing we have observed is that major segments of the public are very concerned about giving commercial organizations access to student data, even if it has, in theory, been anonymized. Because there is no obvious way they might profit from the data financially, there usually is less concern about having researchers within the education institution or in external academic or nonprofit research organizations maintain the data for the life of a research effort.

We expect that more standard practices for data protection and protection of human subjects in keeping with federal guidelines will emerge over the next five years. Indeed, a number of groups, such as the National Academy of Education, have been advocating for more balanced regulations that will make data more available for education research (National Academy of Education, 2017). State laws around data privacy and protection are in much more flux, however, and this situation is likely to continue for some time.

### *Greater Use of Unstructured and Multimodal Data*

Chapters 2 and 3 provided an abbreviated treatment of some of the newer types of data and associated analytic techniques being used to explore learning processes. These include automated systems for tagging video data, automated text mining, and techniques for working with audio file data. While not usually incorporated into today's commercial digital learning products, such capabilities are being used in research prototypes and have intriguing potential for use in improvement research. For example, effective collaboration has been identified as an important skill, and one that many students have yet to acquire. D'Angelo and colleagues (2015) are working to combine audio and learning system data to generate automated indicators of the quality of collaboration among student triads working in classrooms. These investigators have middle school students work together on mathematics problems presented on a laptop while each student wears an individual microphone to pick up his or her speech. A speech activity detection system can be run on each student's audio channel to generate data on who is speaking when. These data are then combined with time-stamped data from the learning system so that each student's speech utterances can be linked to the problem being worked on. Without doing any analysis of the actual content of the students' speech (i.e., what they said), the researchers have been able to generate indices such as equality of participation, which have been shown to correlate with quality of collaboration in prior research (Richey, D'Angelo, Alozie, Bratt, & Shriberg, 2016).

In addition, more advanced techniques like neural networks (recently used to improve Google Translate) are likely to be applied to a greater extent to large, complex education datasets than they have been in the past. These techniques may generate insights heretofore unavailable, but we note that the lessons around the need for collaborative interdisciplinary teams and the need for a systematic, improvement-focused approach will apply just as much to these newer techniques as they do to more widely used analytical techniques.

### *Press for Accountability and Transparency Around Algorithms Used Within Digital Learning Products*

One of the benefits of CDI participation for educators is the opportunity to learn about how students interact with digital learning systems and how to make sense of the data those systems can provide. Many—perhaps most—of these systems are marketed as being "adaptive" and promoting "personalization." Unfortunately, there are no generally accepted definitions for these terms, and marketing materials focus on rosy abstractions (e.g., "giving each student exactly what she needs") rather than explaining concretely how the system responds to students' correct and incorrect answers. As educators become more aware of how digital learning systems actually work, they are going to push technology vendors for more detail. Does the system give every student the same set of assessment items and adapt just by letting students go through the material at their own pace? Does it keep cycling a student who does poorly on an end-of-module assessment through the same quiz items until the student gets 80 percent correct? Does the system diagnose different kinds of errors and give different hints or supports depending on that diagnosis? Or does it just tell the student whether or not his answer was correct? Are there different sets of materials or paths through the materials for students diagnosed as having different kinds of problems? We expect digital learning system providers to face more questions around these issues, and the answers may well more questions decisions.

### *Call for Incorporating Cost Analyses Into Improvement Work*

One of the trends we have seen in our learning technology evaluation work generally is a desire to look at costs and cost savings along with impacts on education outcomes. We believe that this movement will affect CDI efforts as well. In the case of the hypothetical CDI around nursing program completion used as an example previously, analysts would use the estimates of the impact of the early math intervention on program completion in a cost analysis—examining the amount of additional tuition the college received from the students who persisted in the revised program who would have been expected to drop out if they had not experienced the new intervention. These data then would provide a basis for estimating the additional tuition expected if all nursing students receive the math intervention over the next five years. Analysts also would need to look at the cost of the intervention relative to the way the nursing program was structured in the past, and the cost of the CDI effort itself would be part of this analysis. If the CDI effort really has been successful, benefits should far outweigh the costs. Through such analyses, the value added by the effort could be demonstrated in terms that education stakeholders can appreciate.

### *Data Science's Transition to the Post-Hype Phase*

Gartner, a well-known information technology consulting firm, developed and disseminated a framework for thinking about fast-growing new technology applications they call the "Hype Cycle." First, a new technology development serves as a trigger for what they call the "Peak of Inflated Expectations." As people gain more experience with the technology application, they gain a more realistic picture of what it can and cannot do, and the technology moves into the "Trough of Disillusionment." At this point, the technology gets less public attention, but work tends to continue and as the technology becomes more mature and people start using it with realistic expectations, the "Slope of Enlightenment" is entered. Finally, mainstream adoption of the technology takes off in the final portion of the cycle, the "Plateau of Productivity."

A well-known example of movement through this cycle is that of MOOCs, which arguably are now in the Enlightenment portion of the cycle. For big data and data science, though, it seems we are still in the phase of Inflated Expectations. Real solutions to education problems come from hard work, but flashy anecdotes and examples about surprising data patterns are what garner attention. In the next five years, we expect the field of data science to become more transparent about how it works with education and learning data. Likely, big data and data science will lose some of their cachet at this point, but as they move from hype to reality, real change and valuable applications are likely to occur. As data governance and technologies get better and more mature, we expect to see a lull and then a second uptick in interest in applying data science to education once a larger group of people develop expertise and want to pursue this work despite its requirements for attending to myriad details, setting up a data infrastructure, and negotiating with partners with different kinds of expertise. Once educational data science gets past the Trough of Disillusionment and starts to be applied widely to jointly-defined problems of practice cases of improved learning environments are likely to proliferate.

## An Invitation to the Field

We have advocated for a complex, multi-perspective, iterative process for collaborative efforts to leverage education data for improvement. Because the endeavor is challenging and complex, we have offered descriptions of how such partnerships can be realized, so that a team could use this model to move through the work in phases comprising key phases and supporting conditions. But we cannot offer a cookbook for this approach. Instead, we provide enough guidance and scaffolding to inspire and equip other individuals and organizations for the kind of work we have characterized as collaborative data-intensive improvement. It is not our intention to treat our description of the process and the way we have implemented it as inviolable. We recognize that teams are going to have to work hard to build their own partnerships and that it will take courage, creativity, and persistence to sustain and keep on improving their own collective inquiries. We encourage others to elaborate, modify, and reconstruct the practices and tools described in this book based on the specific situations they encounter. This kind of work is both challenging and exciting—iteration, refinement, and knowledge sharing will be key to harnessing data science for educational improvement.

## References

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, *13*(1), 1–14.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. Washington, DC: U.S. Department of Education.

Civitas Learning. (2016). Emerging benchmarks and student success from across the Civitas. Community Insights Report: Issue 2. Retrieved from http://info.civitaslearning.com/hubfs/Community_Insights/issue_2.pdf

Cohen, D., Raudenbush, S., & Ball, D. (2003). Resources, instruction, and research. *Education Evaluation and Policy Analysis, 25*(2), 119-142.

D'Angelo, C., Roschelle, J., Bratt, H., Shriberg, E., Richey, C., Tsiartas, A., & Noyne, A. (2015). Using students' speech to characterize group collaboration quality. In O. Lindwall, P. Häkkinen, T. Koschman, P. Tchounikine, & S. Ludvigsen (Eds.), *Exploring the Material Conditions of Learning: The Computer Supported Collaborative Learning (CSCL) Conference 2015*, Volume 1 (pp. 819–820). Gothenburg, Sweden: The International Society of the Learning Sciences.

Ginsburg, A., & Smith, M. (2016). *Do randomized controlled trials meet the 'gold standard'? A study of the usefulness of RCTs in the What Works Clearinghouse*. American Enterprise Institute. Retrieved September 8, 2016, from www.aei.org/wp-content/uploads/2016/03/Do-randomized-controlled-trials-meet-the-gold-standard.pdf

Penuel, W. R., & Gallagher, D. (2017). *Creating research-practice partnerships in education*. Cambridge, MA: Harvard Education Press.

National Academy of Education. (2017). *Big data in Education: Balancing the benefits of educational research and student privacy*. Washington DC: National Academy of Education.

Richey, C., D'Angelo, C., Alozie, N., Bratt, H., & Shriberg, E. (2016). The SRI Speech-based collaborative learning corpus. In *Proceedings Interspeech 2016* (pp. 1550–1554).

Siemens, G., & Baker, R. S. J. D. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252–254). New York: ACM. doi:10.1145/2330601.2330661

Treaster, J. B. (2017, February 2). Will you graduate? Ask big data. *New York Times*. Retrieved from www.nytimes.com/2017/02/02/education/edlife/will-you-graduate-ask-big-data.html?_r=0

U.S. Department of Education, Office of Educational Technology. (2013). *Expanding evidence approaches for learning in a digital world*. Washington, DC. Retrieved August 8, 2018, from https://eric.ed.gov/?id=ED566873

# Glossary

**A/B testing** Comparing two versions of a web page, application, or other product by showing each version to a random sample of users and measuring their responses.

**Administrative data systems** Systems that store educational data for schools and districts as well as state and federal governments to manage operations and services provided to students.

**Algorithm** A precise set of rules or instructions that outline the computational steps needed to do calculations or solve problems.

**Artificial intelligence** The ability of computers to perform tasks associated with intelligent human behavior, such as reasoning, decision making, and object recognition.

**Attributes** See *Feature*.

**Behavior detectors** A way of using data from digital learning environments as well as sensors and recording devices to infer aspects of human behavior and affect, such as frustration, boredom, and "gaming the system."

**Big data** While imprecise and regularly shifting, big data can be thought of as datasets that require specialized database and software tools to manipulate and analyze.

**Classification algorithm** A type of prediction algorithm that takes inputs, or observations, to predict a categorical known outcome.

**Clustering algorithms** Clustering is a common unsupervised learning method that groups similar observations together. Clustering algorithms differ in the ways of quantifying "closeness" among observations and "differences" between groups of observations. Hierarchical cluster analysis recursively groups similar observations. *K*-means clustering requires a human-specified number of groups with which the algorithm maximizes similarity within clusters and diversity between clusters.

**Cognitive task analysis** A method of understanding and documenting the hidden cognitive activity involved in a task, such as solving a problem or making a decision.

**Confusion matrix** A 2 × 2 table that lists the number of true-negatives, false-negatives, true-positives, and false-positives.

**Continuous improvement** A structured approach for iteratively refining a process, service, or product over time.

**Cross-validation** Used for training a model using a test-train split of the dataset. Cross-validation involves breaking a dataset into a specific number of subsets, holding out one subset, and using the remaining data to train a model that is then tested on the held-out sample. This process happens for each held-out sample and can be repeated a desired number of times.

**Data breach** Access or use of information—sensitive, private, confidential—by unauthorized users. A legal term for an event that requires notification to the affected parties.

**Data fusion** The process of integrating multiple data sources to produce more consistent, accurate, and useful information than that provided by any individual data source.

**Data interoperability** The ability of different information technology systems and software applications to communicate and exchange data in a usable form.

**Data product** An outcome of an analysis in the form of a table, visualization, or algorithm that is intended to communicate something to an audience. Data products can be static, updating, or interactive.

**Data science** The application of data analysis, programming, and domain expertise in order to extract meaningful insights about important issues. Includes the use of statistical approaches and machine learning techniques.

**Data scientist** Person who acquires, manages, and analyzes large complex datasets. Such individuals generally possess a combination of computer science skills, a background in statistics as well as machine learning, and relevant domain expertise.

**Data sprint** An event where data scientists meet up with data providers and other stakeholders to intensively and jointly analyze data over a short period of time, typically less than one week.

**Data use agreements** Written agreements between researchers and schools/districts that outline exactly which administrative data are to be given to the research organization, how the data will be used, and when the data will be destroyed.

**Data visualization** Involves graphically or visually representing one or more features in a dataset so that potential patterns can be discerned by human perception.

**Data warehouses** Digital storage systems that provide access to current and historical data and provide a platform for sharing and exchanging information stored in separate datasets or "silos." A key challenge is to bring together data from systems designed to capture transactions (such as showing up to a particular class on a particular day) with more static data (such as a student's prior courses).

**Database** A software system for updating and accessing data. A relational database stores data in tables. Data from a relational database can be accessed in many different ways based on how data are connected using linking variables.

**Data-driven decision making** Practices for data collection, analysis, and use that support organizational and instructional decision making.

**Data-intensive research** The use of data that stretch the typical storage, computational requirements, and/or complexity that is currently typical of a research field.

**Design research** A system for improving learning environments through iterative cycles of design, development, and implementation. Design research emphasizes collaboration among researchers and practitioners in real-world settings.

**Design-based implementation research** A type of design research in which researchers and practitioners attend to improving learning environments as well as scaling and supporting the sustainability of efforts to improve learning.

**Dimensionality reduction** In statistics and machine learning, refers to ways of reducing the number of features, or variables, in a dataset. Common approaches include principal components and factor analyses.

**Disclosure** Release of private information shared in one context within another context.

**Driver diagram** An improvement tool that outlines a group's theory for how to achieve a desired outcome by identifying necessary and supporting conditions as well as specific change ideas.

**Educational data mining** The use of statistical and machine learning methods to discover patterns in educational data. Often concentrates on identifying patterns within datasets from specific digital learning environments like intelligent tutoring systems; these same technologies typically deliver interventions aimed at improving learning.

**Evidence-based practice** A term used in both healthcare and education to describe an explicit commitment to engaging in practices that are based on the best available research evidence.

**Exploratory data analysis** Can involve some combination of data visualization and feature engineering to understand the structure of and relationships within a dataset.

**Family Educational Rights and Privacy Act (FERPA)** FERPA imposes responsibilities on entities handling education data as a means of protecting parents' and students' rights. It gives parents and eligible students the right to review students' educational records and requires that schools obtain their consent before disclosing any student information contained in those records to third parties—except in the case of a specified set of exemptions.

**Feature** A feature is a column within a dataset that contains data for multiple observations, or rows. Other names for features include variables and predictors.

**Feature engineering** Feature engineering is the process of creating new variables within a dataset using theory-driven, context-driven, or automated approaches.

**Fishbone diagram** An improvement tool used to display factors contributing to a problem or undesirable situation; also called a "cause-and-effect" or Ishikawa diagram.

**Harms** Negative consequences from a privacy breach, such as economic or psychological (e.g., embarrassment).

**Information/data security** Standards for electronic storage (including encryption) of personal information and other data such that only authorized access is allowed.

**Institutional review board (IRB)** Typically committees or individuals who review applications for research and approve/recommend alterations to research procedures based on the legal statutes and ethical requirements governing research with human subjects.

**Intelligent tutoring system (ITS)** Type of digital learning environment that applies artificial intelligence to students' interactions with the system. ITSs collect information on a student, her progress in the system, and interactions that she engages in during a learning task, providing feedback in the form of hints, strategies, and different ways to practice skills.

**Interoperability** See *Data interoperability.*

**Knowledge discovery in databases (KDD)** A general term for the process of extracting information from raw data. The phrase emphasizes that knowledge is the key outcome of any data-driven inquiry.

**Knowledge engineering** Involves using theory and approaches like cognitive task analysis or expert interviews to develop an algorithmic representation of a focal construct.

**Learning analytics** A research field that uses a variety of analytical techniques to guide human judgment, support human action, and optimize learning environments.

**Learning curve** A graphical representation of the relationship between learning trials and quality of performance.

**Learning management system (LMS)** Online system that supports instructors in delivering course content and supporting online learning activities. Within an LMS, instructors and students can share instructional materials, make class announcements, submit and return course assignments, and communicate with one another online.

**Machine learning** An overarching label for data analysis approaches that use supervised and unsupervised algorithms to identify patterns within a dataset.

**Massive Open Online Courses (MOOCs)** Free web-based course designed for delivery to large numbers of geographically dispersed learners. The course is open to anyone with an Internet connection, without consideration of academic qualifications.

**Multimodal analytics** Computational methods for blending multiple data streams from sensors, recording devices, and digital learning environments to identify patterns across data streams.

**Munging** See *Wrangling*.

**Networked improvement community (NIC)** A scientific community that comes together to achieve an improvement goal with a common understanding of the problem they are trying to address and a commitment to trying out new approaches and to sharing knowledge across network participants.

**Open data** Data that are available for anyone to access or use; making open data usable means making it accessible in machine-readable, structured, granular, and well-documented formats.

**Personally Identifiable Information (PII)** Student information that can be used to distinguish or trace one student uniquely through direct or indirect linkage.

**Plan-Do-Study-Act (PDSA) cycle** A method for testing a change idea by developing a plan to test the change, carrying out the test, observing and measuring the consequences, and selecting modifications for the next cycle.

**Predictive modeling** See *Supervised learning*.

**Predictors** See *Feature*.

**Preprocessing** See *Wrangling*.

**Privacy** Control over one's data and an expectation that personal information will only be collected or used within a particular context.

**Regression algorithm** A type of prediction algorithm that take inputs, or observations, to predict a continuous outcome.

**Regressors** See *Feature*.

**Reproducible science** Research carried out and documented in such a way that the data analyses can be duplicated, supported by access to the original data as well as the computational steps taken to process data and generate results.

**Research–practice partnerships** Long-term collaborations between education researchers and educators for the purpose of performing research that can be used to improve education practices and outcomes.

**Sankey diagrams** Illustrate the movement of inputs across key steps, changes, or decision points making up a flow of activity whereby the width of various "flows" is proportional to the quantity of inputs.

**Sensors and recording devices** Physical instruments that are used to capture and store biometric and other data in space and across time, including location, physical movement, and speech.

**Statewide longitudinal data system (SLDS)** A type of administrative data system containing information on all of a U.S. state's public school students and capable of tracking student information over multiple school years and across multiple schools. SLDSs can include a unique statewide identifier for every student as well as each student's demographic characteristics, enrollment history, and scores on statewide accountability tests.

**Structure discovery algorithms** A class of unsupervised learning techniques that identify relationships across features within a dataset without being trained against a known outcome.

**Structured data** Organized for efficient processing. While lacking a precise definition, in general, structured data are any kind of data organized into a table with rows and columns.

**Student information system (SIS)** A type of administrative data system used by schools and universities to store student-level information, such as demographics, course schedules, attendance, grades, and test results.

**Supervised learning** Builds computational models that quantify relationships between features and a known outcome (also called labels or dependent variables). In effect, supervised learning builds a model that can predict a value based on learning from many labeled examples called the training set. This training set, often consisting of human-labeled inputs and outputs, constitutes the supervision of the learning process.

**Unstructured data** In contrast to structured data, these type of data may have an internal structure but do not outwardly conform to traditional ways of organizing data for efficient transactions.

**Unsupervised learning** A way of finding patterns or structure in a dataset without having known outcomes, or labels, to learn a model for the data. Unsupervised learning is useful for understanding hidden relationships among features in one's dataset.

**Variables** See *Feature*.

**Vs, The 4** Volume, velocity, variety, and veracity. Volume is about the amount of data available measured in bytes, a unit of measure in computer memory. Velocity addresses the speed, or rate, at which data are generated. Variety describes the types of data, such as different events tracked within a digital learning environment or the variety of data types used in an analysis, such as audio and video. Veracity captures the degree to which one can trust data. There are no standard units of measure for veracity, but data can be untrustworthy for a variety of reasons, such as data entry errors.

**Wrangling data** Refers to the work of manipulating and cleaning data, and includes identifying, acquiring, and importing data into analysis software.

# Index