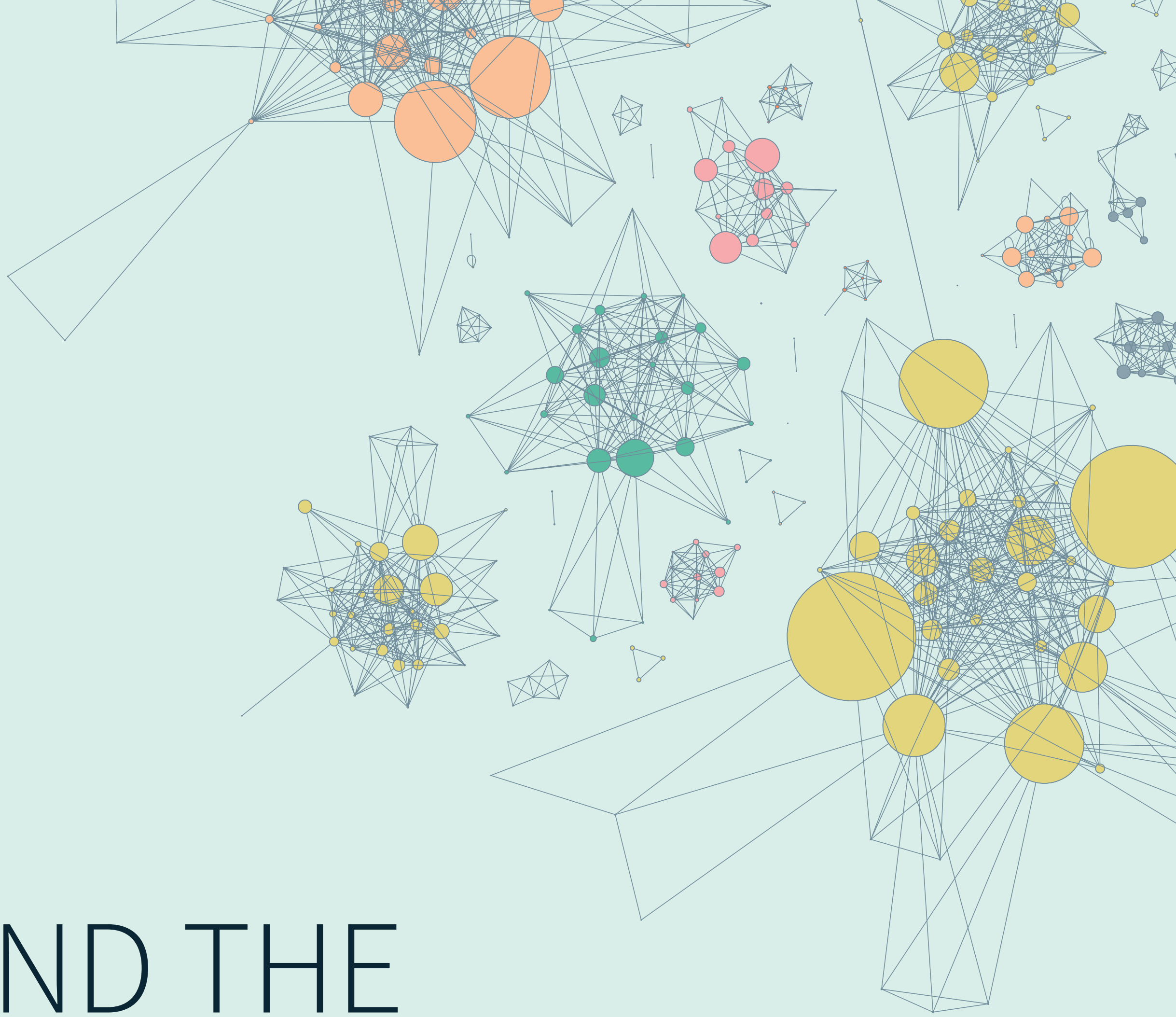


BEYOND THE HYPE OF BIG DATA IN EDUCATION

Practical lessons and illustrative examples of how to derive reliable insights in learning analytics

Rasil Warnakulasooriya // Adam Black

macmillan learning

CONTENTS

—

AUTHORS

Rasil Warnakulasooriya & **Adam Black**

with Contributions from

Florin Bocaneala
William Galen
Jon Harmon
Aida Kavara
Daniel Lawrence

—

ACKNOWLEDGEMENTS

We would like to thank the leadership of the learning science and insights team—Dr. Jeff Bergin and Dr. Kara McWilliams. We also thank Dr. Elana Zeide for her comments and counsel on the draft.

Preface

by Ken Michaels, CEO, Macmillan Learning

We are living in challenging and exciting times in education.

Challenging because instructors and institutions are under pressure to improve outcomes, technologies are evolving at a dizzying pace, and costs are rising for students, yet more are struggling or failing.

Exciting, because we are learning more than ever through research about how students study, behave, and learn and how to help them to be more successful; about teaching and institutional excellence of ‘what works’; and about how to facilitate improvements with innovative programs and technologies.

A firehose feeding this research is the volume and granularity of data from the increasingly sophisticated and comprehensive digital solutions that students, instructors, and institutions are using – from teaching, learning, and assessment, through attendance, in-class polling, and surveys, to course and career planning and much more.

Given the wealth of data, we should be drowning in insights, right? Not necessarily. Learning is highly personal, and education is a complex ecosystem. So, sifting out reliable and practical insights from the morass of data is a skilled, multi-disciplinary endeavour. We know from decades of experience that it’s easy to draw conclusions, but it’s much harder to draw reliable insights. That’s what this White Paper is all about – moving beyond the hype about “big data” and getting down to practical tips that will make anyone analyzing data in education more successful.

Our company mission is to improve student success. We achieve this through our deep and collaborative partnerships with the instructors and institutions we serve. It’s because of our mission and these relationships that we have chosen to share the hard-won best practices developed by our Learning Science and Insights teams of analysts and reveal examples of the surprising and practical insights these can reveal.

We hope you find this White Paper interesting and that it contributes constructively to the broader discussion of how data can be used in education responsibly and effectively to improve student success. Have fun reading!

Sincerely,
Ken Michaels

Introduction

The educational landscape today is being rapidly transformed due to innovation in technology and its adoption in teaching and learning. Given that the technologies adopted are largely digital, volumes of data are now available at high velocity and with greater variety for educators and educational service providers to support teaching and learning. The availability of this volume, velocity, and variety of data generated in educational contexts, however, does not mean that relevant insights can be readily and easily be drawn to facilitate greater outcomes for learners and instructors. What, then, is the promise of educational data and learning analytics?

In this paper, we outline the approach to educational data mining and learning analytics we have developed at Macmillan Learning, illustrate how this approach is aiding the development of a next generation of digital educational tools, informing textbook authors and content developers, and enabling better insights and intervention strategies for institutions, faculty and staff, and students themselves. But first, we will begin with a brief survey of the current state of analytics.

The Current Landscape of Educational Analytics

> We believe... that data analytics has a much larger role to play in support of teaching and learning, including understanding of learning processes themselves.

In today's world, analytics has been leveraged mostly in retail businesses. According to a report by McKinsey & Company (McKinsey Global Institute, 2016), even manufacturing, healthcare, and the public sector are underutilizing analytics. The report did not identify education as an area where analytics can play a significant role except as part of personalization. At Macmillan Learning, we take a different view on the value of analytics in support of teaching and learning. We believe that personalization is only one aspect where analytics can be leveraged and that data analytics has a much larger role to play in support of teaching and learning, including understanding of learning processes themselves. Insights from learning analytics can help guide the development of more engaging and effective learning products and courses; the creation of content that is pedagogically rich and sound for instruction, learning, and assessment; provide institutions, instructors, and learners with insight into performance, behaviors, and interventions; support and complement research on impact of learning products (McWilliams et al., 2017); and to shed light on fundamental questions in teaching and learning, such as knowledge transfer.

Professionals with a variety of skills occupy the current data analytics landscape. These skills tend to fall into three categories: mathematics and statistics, domain expertise, and computer programming (Hayes, 2014). Some professionals may have skills in more than one category. To conduct analyses that support education, the ideal combination of skills—that of a data scientist—is the possession of all three, so that relevant analyses can be conducted to draw sound inferences. This combination of skills is at the heart of analytics at Macmillan Learning and is nurtured to derive robust and meaningful insights from data.

Data analysts use a wide spectrum of statistical practices: these range from simple point statistics - such as averages and medians - to sophisticated machine learning techniques. Analyses can be augmented by outlier detection methods, linear and non-linear regression techniques, application of probability models, and classification and prediction models as necessary. Machine learning approaches - which include regression, classification, and prediction techniques - are being used more and more due to the increasing volume, velocity, and variety of data, as well as advances in computation, processing power, and data storage. (Analysts use the term *volume* to refer to the amount of data available, the term *velocity* to refer to the speed at which data becomes available, and the term *variety* to refer to the diversity of data.) Other recent advances in machine learning have taken place in the realm of deep learning, in which added layers of “neurons” along with new algorithms (such as reinforcement) enable computers to “learn” from vast amounts of data, recognize patterns and apply them to specific complex settings, such as playing the game Go with humans (Silver et al., 2016). At Macmillan Learning, we believe that using analytics to draw insights from data is inseparable from sound statistical practice.

The tools that support data analytics cannot be separated from the analytical methodologies. The practice of data analytics is enhanced by open-source analysis tools and statistical programming languages, such as R, which are readily updated with new analytical and visualization packages. Open-source data frameworks (such as Hadoop), data-processing tools (such as Spark), and computation libraries (such as TensorFlow) facilitate the handling and computations of large volumes of data. The ultimate value these tools bring to analytics depends on their end use and the significance of the problems they solve.

> “The availability of this volume, velocity, and variety of data generated in educational contexts, however, does not mean that relevant insights can be readily and easily be drawn.

While machine learning techniques—and more broadly, artificial intelligence (AI)—hold great promise in areas such as speech and image recognition, text reading, language translation, navigation, healthcare, and games, their application in education is currently fairly limited and includes predicting grades, channeling interventions by instructors, administrators, and guidance counselors, and indexing large volumes of educational content. We believe that despite the hype, many challenges remain in applying machine learning and AI at scale in education. Those challenges include, but are not limited to:

1. Educational settings are highly variable

Educational settings are highly variable, which means there isn't enough data within a given context for machine learning algorithms to be reliably trained on. For example, information about the context within which teaching and learning is taking place is largely absent from current data sets. (Context here means, for example, where, and under what conditions teaching and learning are taking place, and for what goals.) In addition, instructors teach students who are in various stages of learning and utilize pedagogical approaches that may differ from their peers.

> Despite the hype, many challenges remain in applying machine learning and AI at scale in education.

2. A large gap exists in the data captured between a task being presented to a learner and its completion

A large gap exists in the data that occurs in the time between a question or a task is presented to a learner and its completion and the part of their work a digital learning and assessment platform can capture. This lack of information—for example, intermediate steps taken before arriving at a solution to a question, classroom events, discussions with a tutor, or study sessions with peers—impacts an algorithm's ability to emulate a student's learning or problem-solving process.

3. Varied uses of content and use cases

Educators vary in their use of content and how and when they choose to use them. These result in further diversity of use cases and contexts.

4. Conceptual domains vary

Conceptual domains vary, which further limits the availability of sufficient data. (For our purposes, the term conceptual domains means distinct disciplines—such as English composition, Economics, or Physics—as well as conceptual differences that may be present within a given discipline.)

It may be possible to constrain the application of machine learning and AI techniques within specific domains, but probably not at scale across domains in the foreseeable future.

Almost all these challenges are, fundamentally, data limitations, which may be overcome in the short term within narrowly defined contexts of teaching and learning in specific disciplines. However, we believe that not being able to support a wide range of educational contexts and disciplines is a severe limitation of machine learning and AI techniques. Given that there is historical precedent to overestimate the impact of technological innovation in the short term and underestimate their influence in the long term (Brooks, 2017), we are cautiously optimistic about the application of machine learning and AI techniques at scale in the domain of education and believe that their potential has only just begun to be widely researched.

Learning Analytics: Recommended Best Practices

At Macmillan Learning, we believe that the guiding light for educational data mining and learning analytics should be the understanding of and empathy for what learners, instructors, and other stakeholders in education are trying to achieve. This steers data analysts to craft solutions that facilitate the needs of learners, instructors, and other users, factoring in relevant research and insights from the learning sciences, impact research, user experience, and user-centered design. This approach also reframes the role of data scientists to that of problem-solvers, freeing them from methodological constraints, and allowing them to innovatively conduct fit-for-purpose analyses in any problem space in the complex arena of education.

> **The guiding light for educational data mining and learning analytics should be the understanding of and empathy for what learners and instructors are trying to achieve in their learning and teaching.**

This approach naturally positions an analyst to navigate through noisy data (in the sense that the contexts within which data originates are not always clear, and raw data does not necessarily encode useful insights directly), analyze data from multiple angles, and reach reasonable conclusions within the data constraints, statistical and methodological limitations, and uncertainties, and yet serve the needs of learners, instructors, and other stakeholders.

"LEARNING ANALYTICS **is both a science and an art.** The science provides a path to interrogate the data and reach logical conclusions, however complex that path may be. The art is to know what to look for in inherently noisy data."

Learning Analytics is both a science and an art. The science provides a path to interrogate the data and reach logical conclusions, however complex that path may be. The art is to know what to look for in inherently noisy data and reach reasonable conclusions that are insightful and actionable amid statistical and methodological uncertainties. At Macmillan Learning, we utilize both in conducting learning analytics. In treating learning analytics as both a science and an art, and from decades of experience, we have developed a set of guiding principles that we hope may be of value to other data analysts in education. We recommend that data analysts:

1

Explore data from multiple angles, and mine data conditionally to find the most appropriate form of analysis that fits the purpose. Conditional data mining offers an integral path for an analyst to explore various hypotheses and to test them efficiently.

2

Visualize data at every step during exploration as much as possible. Data visualization can help avoid premature or unwarranted conclusions. For example, avoid reporting correlation values from linear regression before visually inspecting if the plot shows a linear trend.

3

Avoid foregone conclusions: that is, avoid looking for data to support preconceived conclusions. Instead, make conclusions that stand out from the data and within the assumptions and constraints of the analysis.

4

Understand how users engage with learning applications before conducting analyses. Avoid treating applications as black boxes and data as context-free. User experience and educational impact research can provide essential context of users and supplement or complement the platform data available to analysts.

5

Do not reject outlier data without due consideration. Outliers often lead to surprising insights. In turn, these can inform much broader conclusions.

6

Use point statistics cautiously, as the blind use of them can often mask the “signal” from the “noise”. In other words, the “median isn’t the message” (Gould, 1991).

7

Isolate and adjust for confounding factors within data as much as possible. Confounding factors pose significant challenges in analytics. Forming hypotheses about these factors and analyzing their effects can help tease out their relative influence on the observed data.

8

Safeguard against overfitting statistical models. One goal of data analytics is to generalize findings from a specific data set into the future. Statistical models developed to explain a current data set extremely well (that is, the model overfits the data) tend to fail to explain future data well. Instead, develop statistical models that balance fit and generalizability.

9

Strive to understand predictive and machine learning algorithms and their performance in a wide variety of educational circumstances. Never treat these algorithms as black boxes with data as mere inputs and results as mere outputs. Analysts working in education have a responsibility to avoid biases (explicit or hidden) and to be careful with predictions and recommendations.

10

Strive to obtain quality data from product applications because improving student and education outcomes depends on it. Platforms should be instrumented so the data they capture reflects user engagement with enough granularity for the research questions at hand.

11

Align statistical thinking with scientific thinking. For example, look for repeated patterns in the data and for the educational and practical significance of the findings rather than relying on statistical significance tests alone.

12

Consider how learning analytics can contribute to foundational empirical research in teaching and learning instead of limiting its use to simply summarizing observations, testing hypotheses, or developing algorithms (predictive or otherwise).

These principles guide data analysts at Macmillan Learning so that their research provides reliable and actionable insights for the development of effective learning solutions and educator support.

Learning Analytics: Illuminating Examples

Macmillan Learning leverages data to continuously improve its educational products and content and to develop new learning platforms and media that facilitate greater outcomes for students, instructors, and other stakeholders. In what follows, we share a number of illuminating examples that practically demonstrate the journey toward obtaining valuable insights, examples of how to avoid erroneous conclusions, and findings that we hope are of interest to educators and educational data analysts.

What Types of Content Impact Learning the Most? Beware of First Impressions

This example shows how data mining can inform product development and how actionable insights may be best derived by analyzing data from multiple angles.

The digital product used in this example assesses foundational skills and guides remediation. First, a student takes a pre-test for a subject area. Based on the student's performance on the pre-test, they are provided a personalized study plan, which is followed by a post-test. The test performance data and the study plan engagement data provide instructors with insight into the foundational skills and gaps between each student's knowledge.

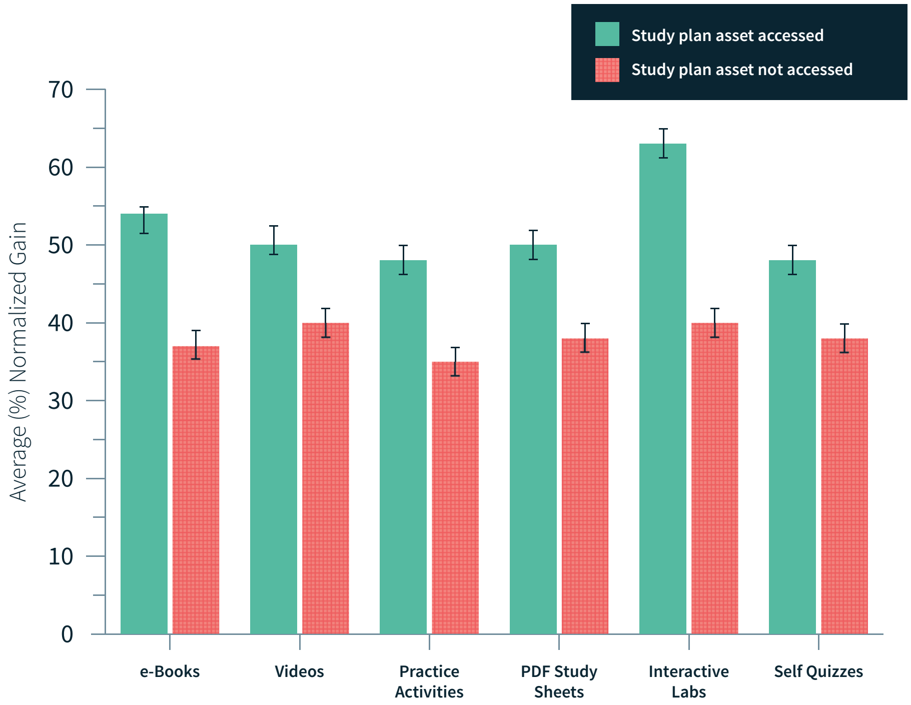


FIGURE 1. Students who accessed the study plan show larger gains from pre- to post-test. At first glance, interactive labs seem to have the biggest impact on student gains. (Statistical uncertainties shown are standard errors.)

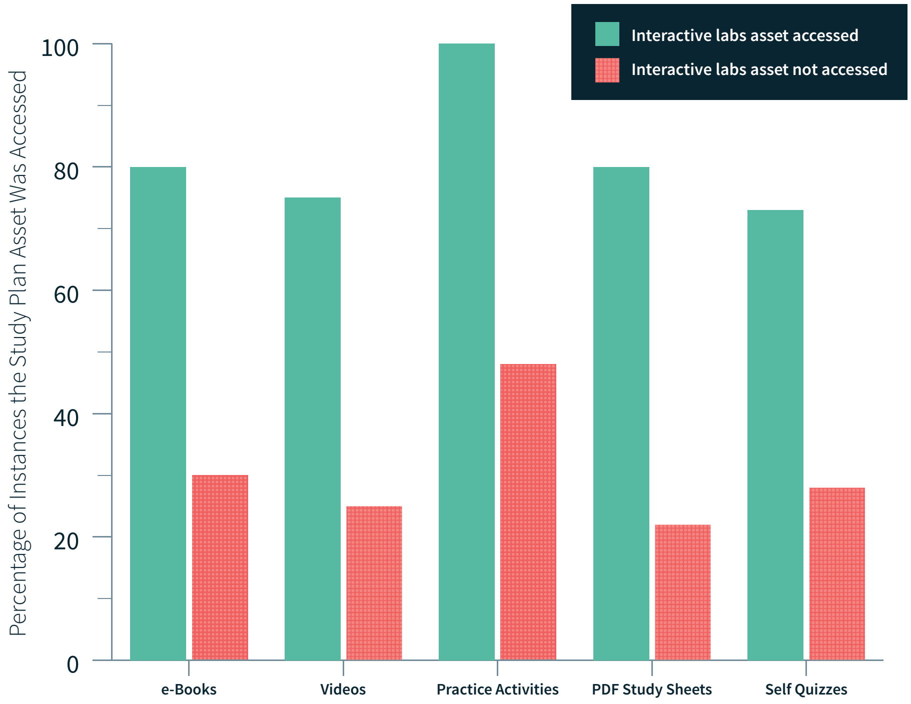


FIGURE 2. A closer look at the data shows that students who accessed interactive labs also frequently accessed practice activities, followed by the eBook and self quizzes. Thus, the results seen for interactive labs in Figure 1 must be rethought.

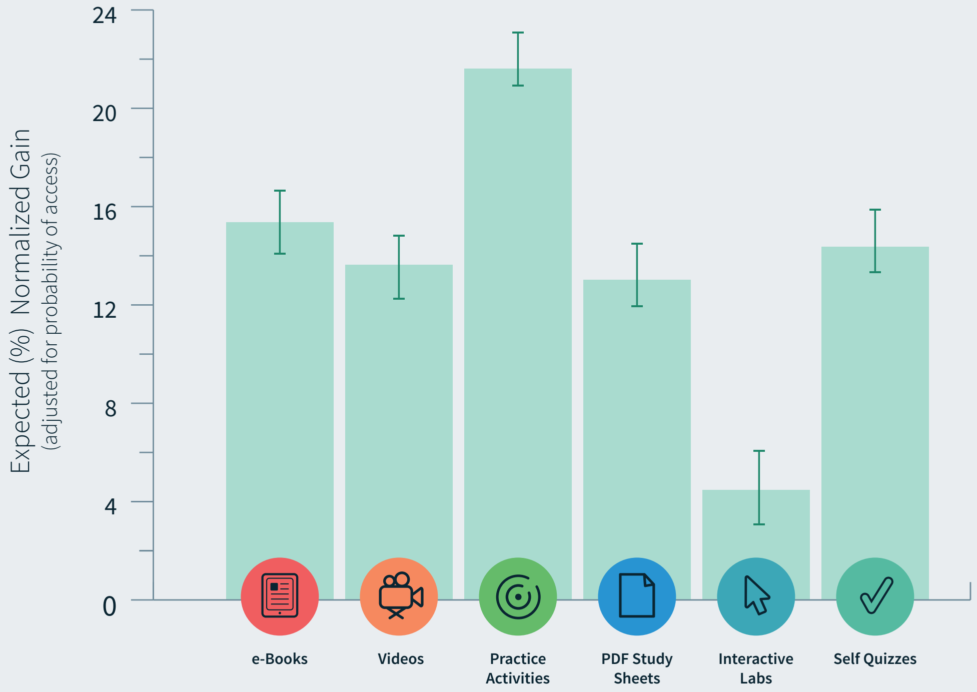


FIGURE 3. Adjusting the data in Figures 1 and 2 conditionally and probabilistically, the most reasonable conclusion is that practice activities—not interactive labs—have the biggest impact on student gains. (Statistical uncertainties shown are standard errors.)

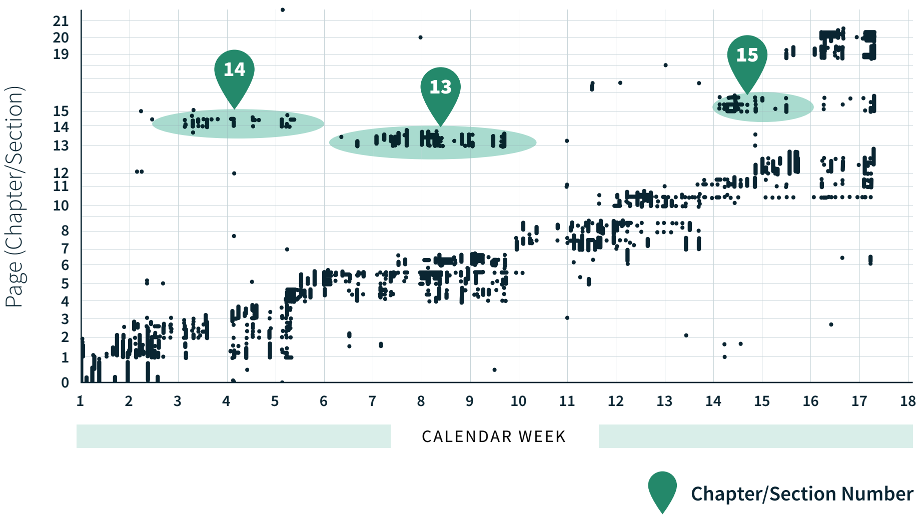
A comprehensive analysis of data from pilot users revealed that among the various assets available in the study plan, the most effective are likely to be the practice activities and self quizzes. However, an initial analysis suggested quite a different conclusion - that interactive labs were the most effective. It was only by looking at the data from several perspectives, mining data conditionally, and adjusting for conflating factors probabilistically that practice activities and self quizzes emerged as the content with a higher likelihood of impacting students as determined by the pre- to post-test score gains. This particular example illustrates (see Figures 1–3) several key aspects of Macmillan Learning's approach to educational data mining and learning analytics: that of taking a multidimensional and conditional approach to learning analytics and maintaining a sensitivity to confounding factors in order to extract cautious and reasonable conclusions from data.

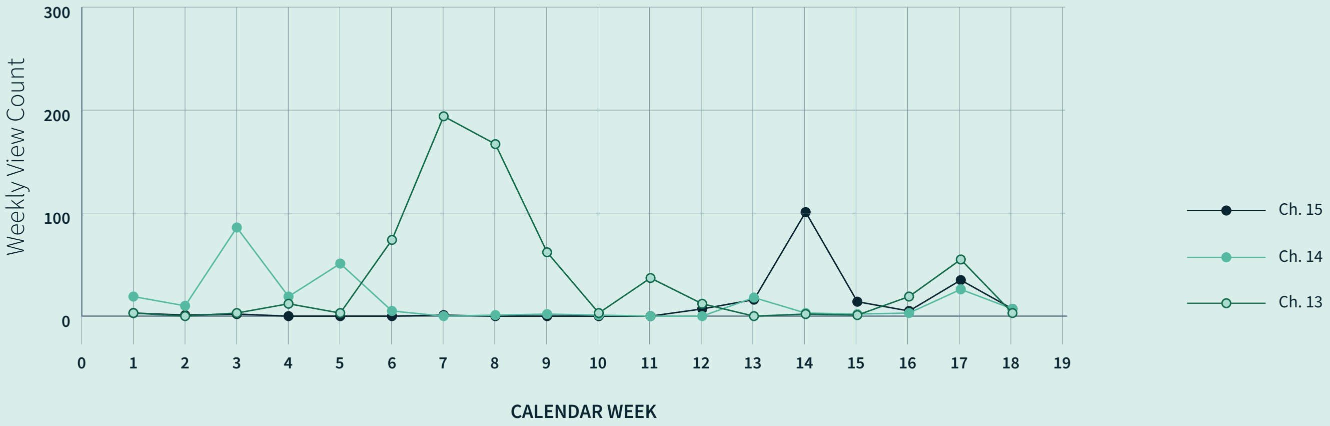
How Can Learning and Assessment Content Be Effectively Improved? Insights into eBooks and Learning Activities

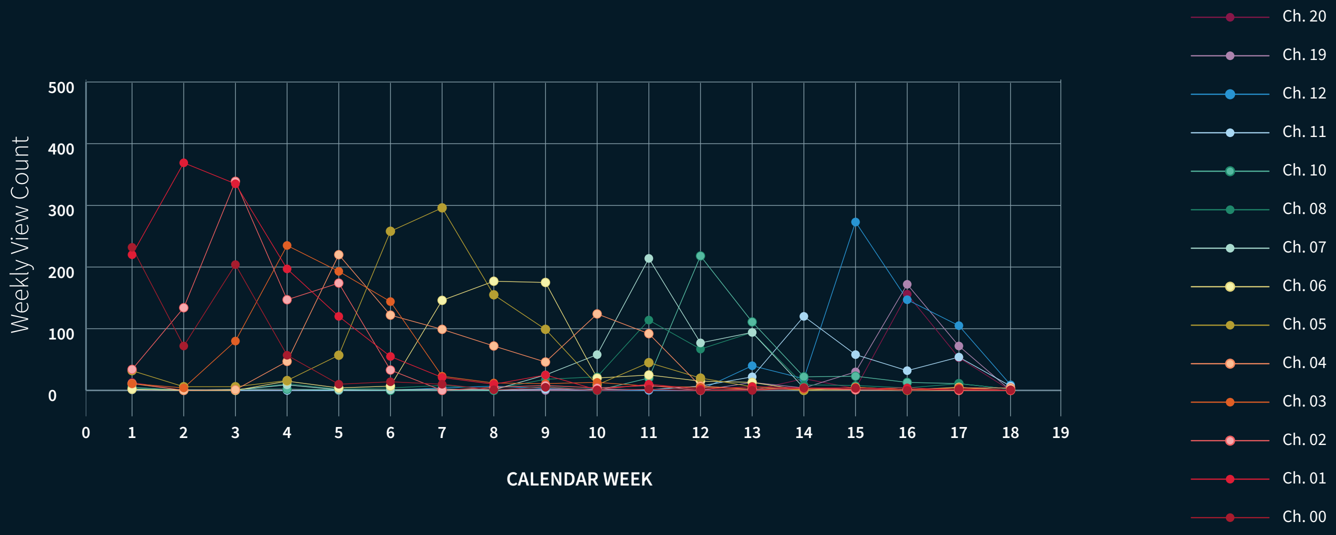
Effective outcomes cannot be expected for instructors and students unless they have access to effective content to teach and learn from. Macmillan Learning uses a wide range of analytics techniques to create effective data-based feedback loops between instructors and students and authors, editors, and content teams. The content usage and quality metrics obtained from data analytics help guide the improvement of textbooks, both print and digital, and continuous empirical refinement of online assessment and learning content.

For example, the analysis of eBook usage (see Figure 4) provides insights into course design in terms of scope and sequence. By analyzing student usage of eBook content—including how much time students spend on pages, how often students access content over time, and how these behaviors and preferences change during the course—we can optimize the scope, sequence, coverage of content, and functional tools that aid studying.

FIGURE 4. Usage of eBook pages (chapters and sections) over calendar weeks of a course. The top chart reveals that most chapters and sections are covered in the order listed in the book's table of contents, but some are covered out of sequence and earlier than expected (e.g., chapters 13, 14, and 15). The middle and the bottom charts show the chapter and section page view trends over calendar weeks for chapters 13, 14, and 15, and the rest of the chapters covered in the course, respectively.



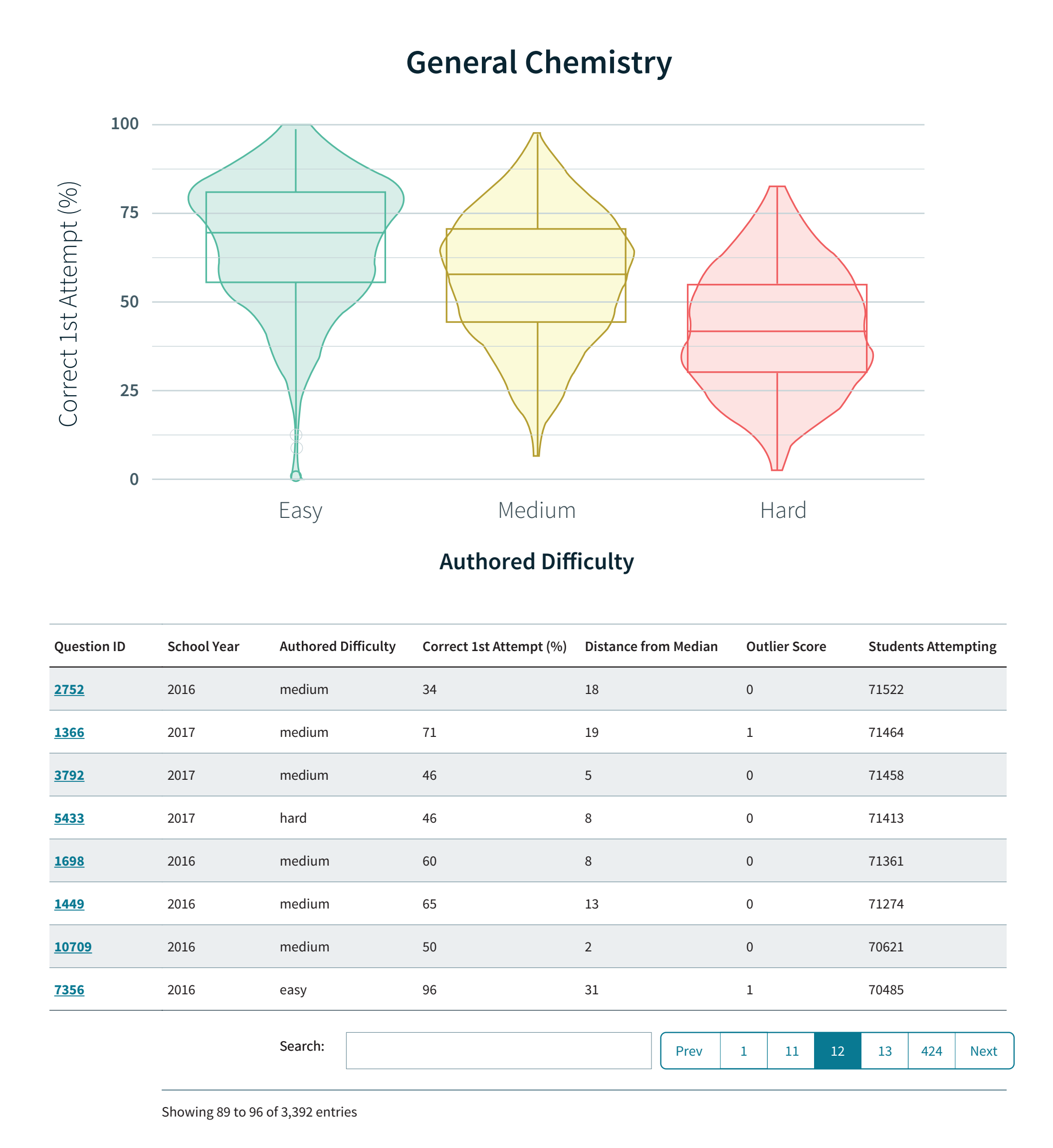




Learning Analytics: Illuminating Examples

Data mining can also be used to analyze the usage and performance of learning and assessment items. This enables efficient, empirical feedback loops—between the students and instructors using the content, and the authors and teams refining the content and functionality—that guide continuous improvement. This has led to the creation of at-scale internal analytics solutions and architectures that facilitate continuous improvement of content (see Figure 5). However, making these feedback loops actionable is often challenging due to noisy data and the variety of user contexts and use cases. Novel analytical techniques are being developed at Macmillan Learning to overcome these challenges.

FIGURE 5. A dashboard displaying content performance and usage that enables authors and content teams to efficiently and effectively improve products. This screen displays the distribution of the difficulty of learning activities (as measured by a correct-on-first-attempt metric) compared with the author-expected difficulty (green = easy, yellow = medium, red = difficult). This enables internal users to assess where the content programs are performing as designed and where opportunities for improvement exist.



Question ID	School Year	Authored Difficulty	Correct 1st Attempt (%)	Distance from Median	Outlier Score	Students Attempting
2752	2016	medium	34	18	0	71522
1366	2017	medium	71	19	1	71464
3792	2017	medium	46	5	0	71458
5433	2017	hard	46	8	0	71413
1698	2016	medium	60	8	0	71361
1449	2016	medium	65	13	0	71274
10709	2016	medium	50	2	0	70621
7356	2016	easy	96	31	1	70485

Showing 89 to 96 of 3,392 entries

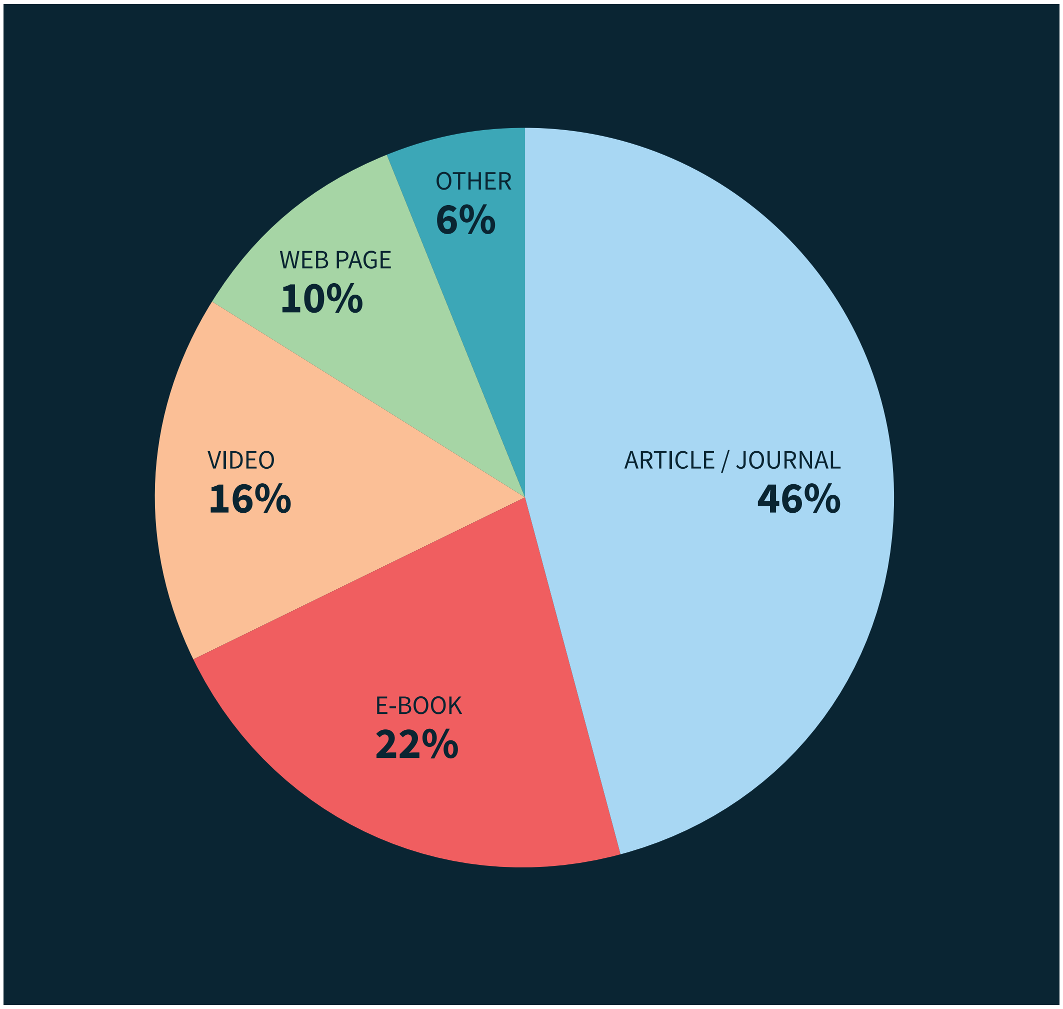


FIGURE 6. Usage of all instructional and assessment content at a sample institution, including locally subscribed and open educational resources (OER). In this analysis, instructors favored content such as journal articles and eBooks over videos and web pages.

Open Educational Resources: Exploring Use and Enhancing Search Using Natural Language Processing

In the current educational landscape, there is growing interest in using open educational resources (OER) for courses, and supplementing publisher-provided content with OER. Thus, OER have become part of the instructional learning equation. However, the usage and effectiveness of OER remain largely unexplored due to a number of fundamental challenges (Seaman & Seaman, 2017). These challenges vary considerably, and include discoverability (e.g., lack of a central index), content sufficiency in a subject domain, recency (content being stale), sufficiency of content quality, diversity of use cases, and variability in the rights of digital content. These varied challenges make the curation of OER content, data collection, and content improvement challenging. Macmillan Learning has introduced solutions to address many of these barriers by aggregating OER at scale, indexing them against course learning outcomes, and measuring usage and user engagement (see Figure 6). In addition, we are further attempting to measure the quality and effectiveness of OER (e.g., by assessing the impact of subsequent outcomes for students and instructors).

In the near future, insights into the usage and effectiveness of open educational resources can help educational institutions better understand how to optimally achieve outcomes with available materials. For Macmillan Learning, these insights allow for the continued support of instructors, institutions, and students in their efforts to achieve the best outcomes possible with a range of educational tools and content. These insights can contribute to a discourse about how to create the most pedagogically effective OER and empirically improve them for use by communities of educators.

In order to identify ways to enable instructors to more effectively utilize OER in their courses, analysts at Macmillan Learning have applied natural language processing to automatically categorize and index more than 6 million OER assets (at the time of this writing). This machine-learning driven taxonomy enables instructors to efficiently discover and select OER assets by topic, type, and/or course learning outcome.

> These insights can contribute to a discourse about how to create the most pedagogically effective OER and empirically improve them for use by communities of educators.

Supporting Learning Objectives-Driven Instruction Through Data Analytics

Calls for accountability in higher education and the changing landscape of college learners have resulted in an increased focus on high-quality, evidence-based teaching practices in the classroom (Gyurko et al., 2016). A wealth of research suggests clear, appropriately challenging learning objectives are a foundation for effective course design (Mayer, 2008) and have the ability to positively impact student learning outcomes (Hattie, 2009). Combined with student engagement and performance data, new opportunities are emerging to validate instruction driven by learning objectives at a scale that have not been possible in the past. Understanding how learning objectives are used in courses provides a first step toward supporting instruction that is based on learning objectives (see Figure 7). By measuring student performance against learning objectives, instructors and researchers can develop an understanding of how instruction is enabling students meet the desired objectives in a course. Such understandings also provide a foundation for personalization and adaptive learning, for timely and successful intervention strategies, leading, in turn, to the generation of new pedagogical models.

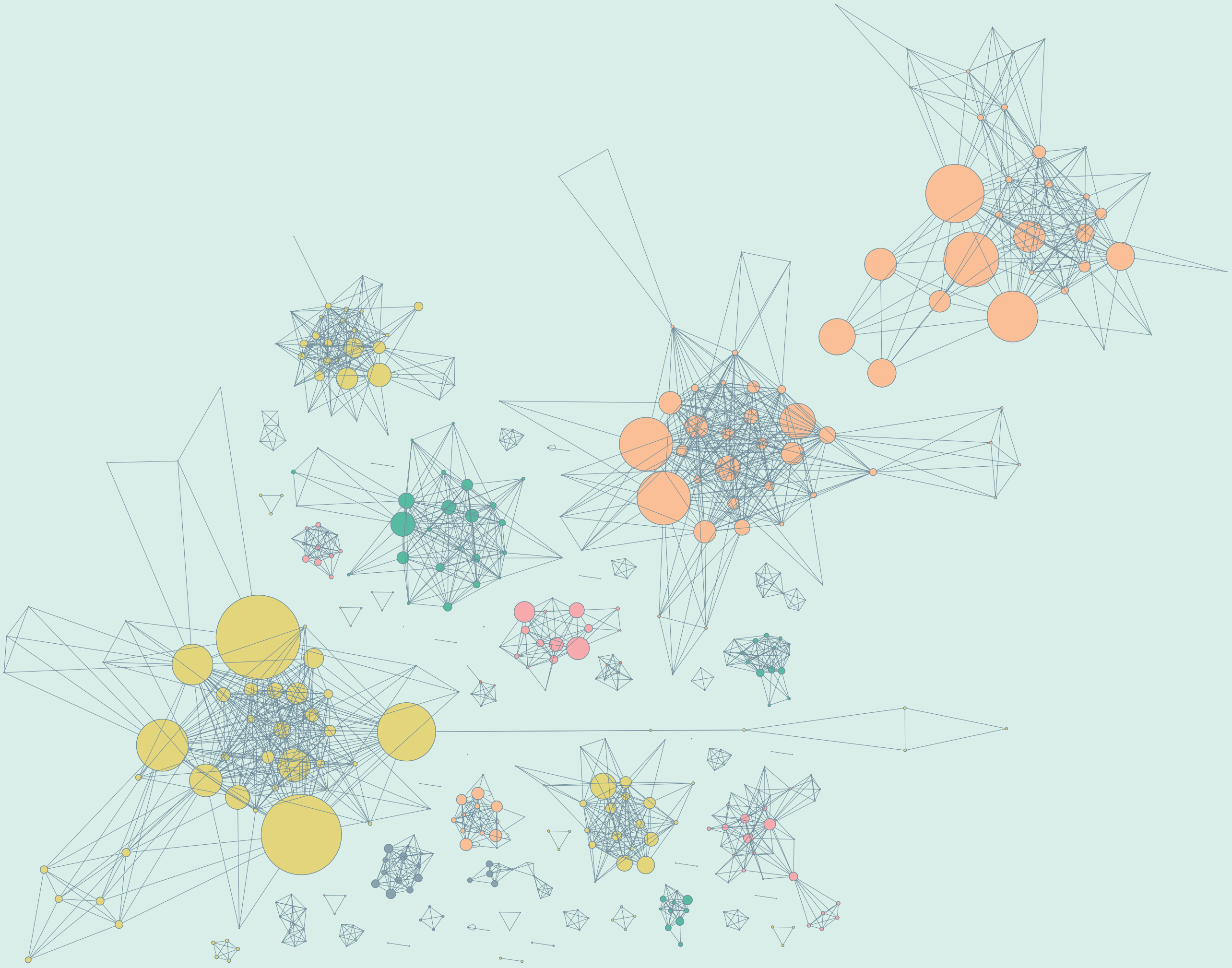


FIGURE 7. The network of learning objectives used in English composition courses. Each node (circle) represents a learning objective. The more assignments a learning objective was included in, the larger the node. Learning objectives appearing in the same assignments are linked, with the closeness of the nodes showing how frequently a pair of objectives are included in the same assignments. These analyses have the potential to support instructors' needs in learning objectives-driven courses.

Learning Analytics: Illuminating Examples

Active Learning: Empirical Insights into a Growing Pedagogy

Active learning strategies continue to be used by instructors, and a growing body of educational research and evidence demonstrates their ability to improve a range of student outcomes. Educators implement active learning in a variety of ways, from interspersing traditional lectures with polling to flipped classrooms (Freeman, 2014).

Macmillan Learning provides a widely used in-class polling system. This enables insights at scale on trends in pedagogical approaches per discipline and class size, as well as insights on students' in-class participation (see Figure 8). These data further provide insights for product innovation, enabling recommendations on best practices and effective in-class interventions in active-learning contexts.

> This enables at-scale research into active learning pedagogies by discipline and class size, detailed insights into students' in-class engagement, and relationships with performance and retention.

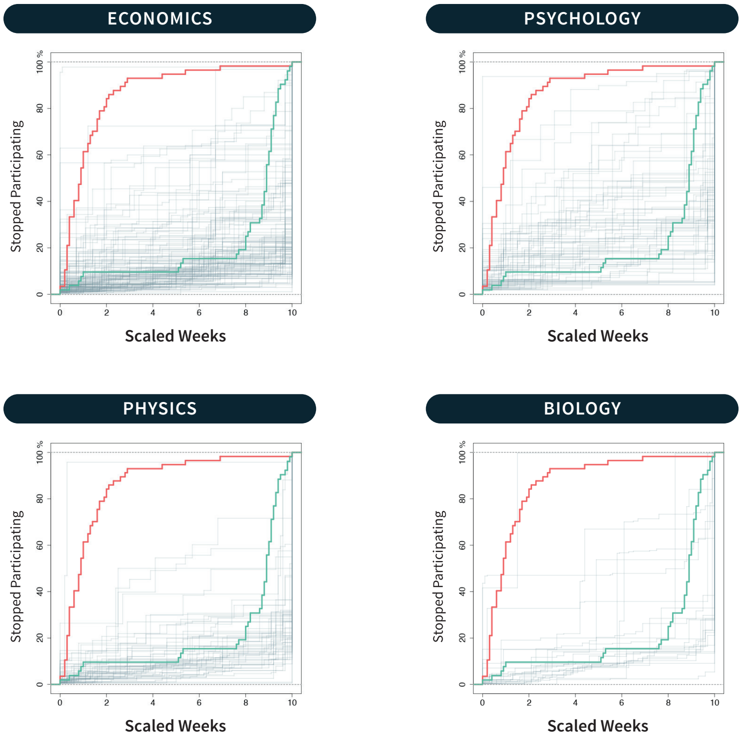


FIGURE 8. How student participation varies over the duration of courses and for different disciplines measured using an in-class polling system. Each grey track within a discipline graph shows the progression of an instructor's course, week by week (from left to right), and what fraction of their class stopped participating (from bottom to top). Courses tracking close to the red trend have students disengaging early on in the course. In contrast, courses tracking close to the green trend have students engaging throughout the course. (Sharp vertical jumps at right reflect the natural wrap up of courses when students tend to stop participating.)

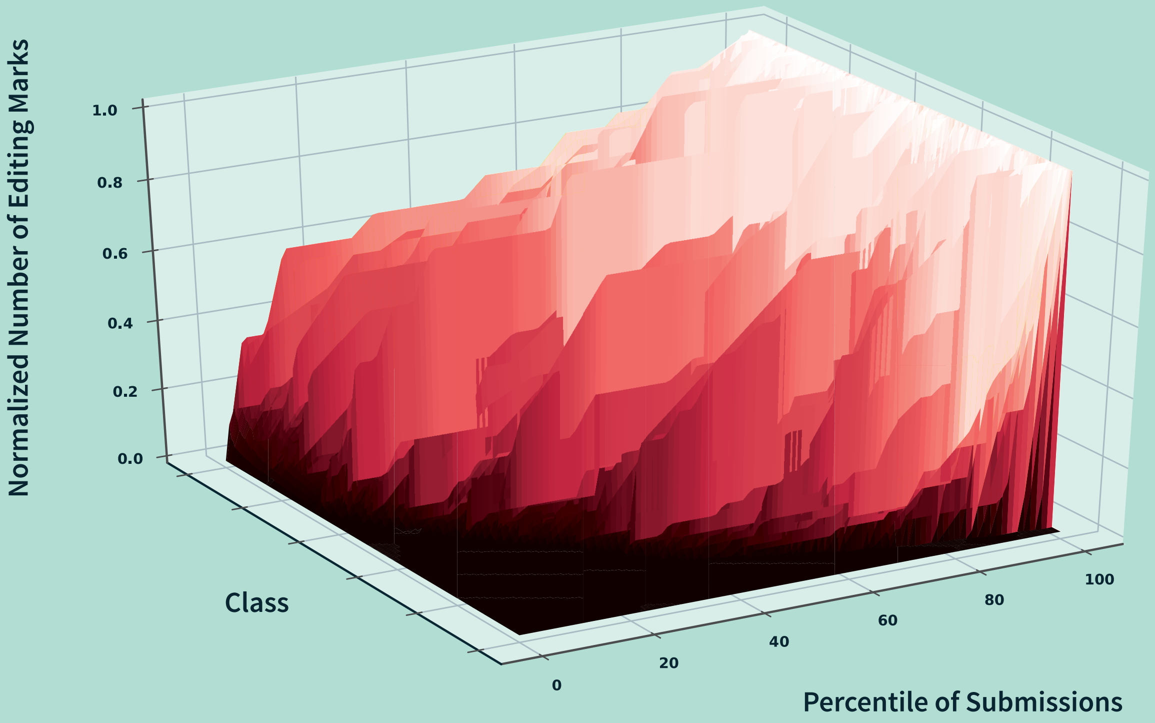


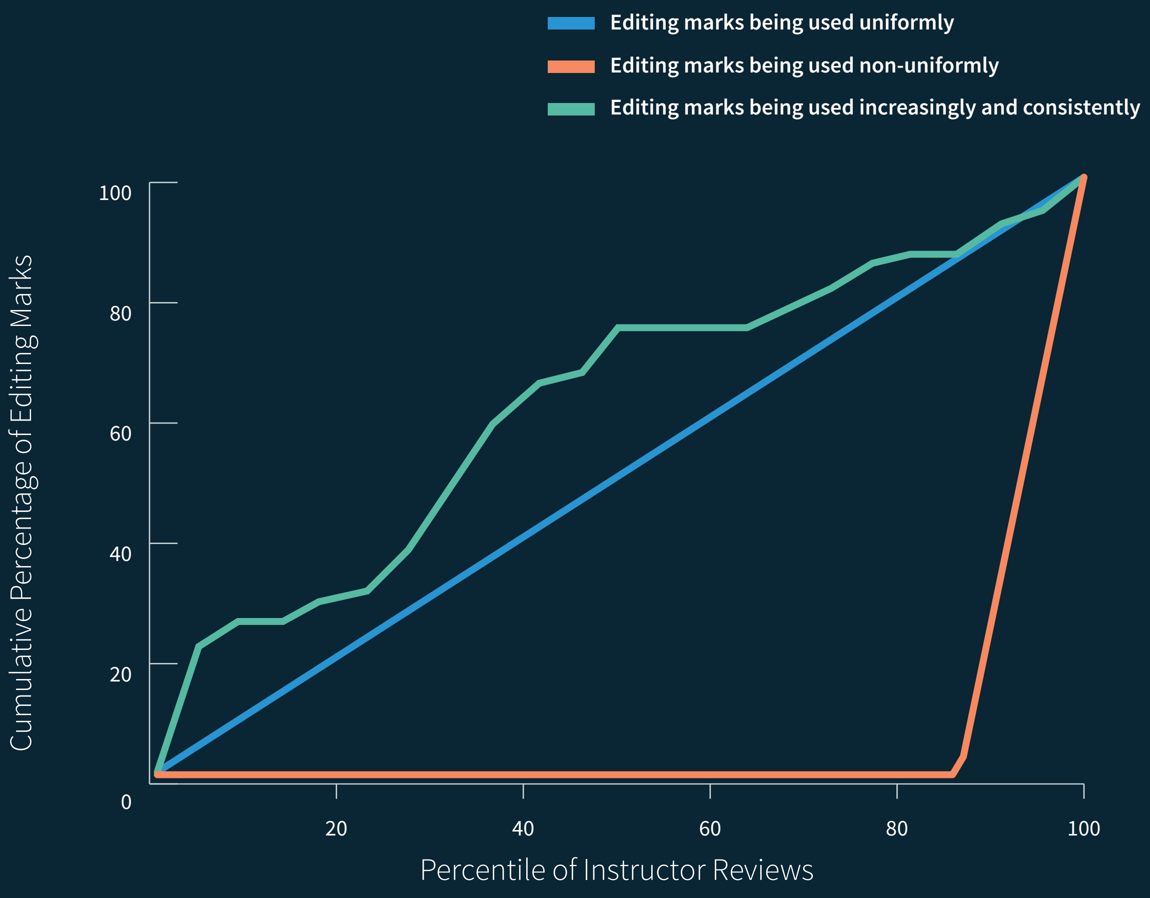
FIGURE 9a. Instructors use editing marks on student compositions either uniformly across the entire class (evident by trends that ascend consistently from left to right) or non-uniformly (evident by trends that stay flattened toward the right "wall" of the graph and then ascend sharply). The latter is the most common usage of editing marks.

How Do Instructors Provide Feedback on Student Compositions? Extracting Insights from Noisy Usage Patterns

This example demonstrates how careful analysis can reveal distinct segments of users, despite noisy and complex data.

In an English composition product offered by Macmillan Learning, an analysis of instructor usage revealed two distinct approaches to providing feedback on student compositions: instructors who provide summative comments to a student's overall composition and instructors who insert specific feedback directly within the student's composition using editing marks. Furthermore, the latter group of instructors use the editing marks tool in two distinct ways: those who use the editing marks uniformly across all their students' compositions (i.e., more or less an equal amount of editing marks per composition) or those who use the editing marks non-uniformly, which is more common (see Figures 9a and 9b). These observations of instructor preferences guide product development and refinement to ensure that the desired editing and feedback tools are available in most effective formats.

FIGURE 9b. A slice through Figure 9a reveals the two main approaches instructors use to provide feedback on student compositions: Instructors use editing marks on student compositions either more or less uniformly across the entire class (green) or non-uniformly (orange). Each curve corresponds to a different course pattern. The blue line represents a hypothetical "ideal" course in which an instructor provides a consistent number of editing marks for every student and composition.



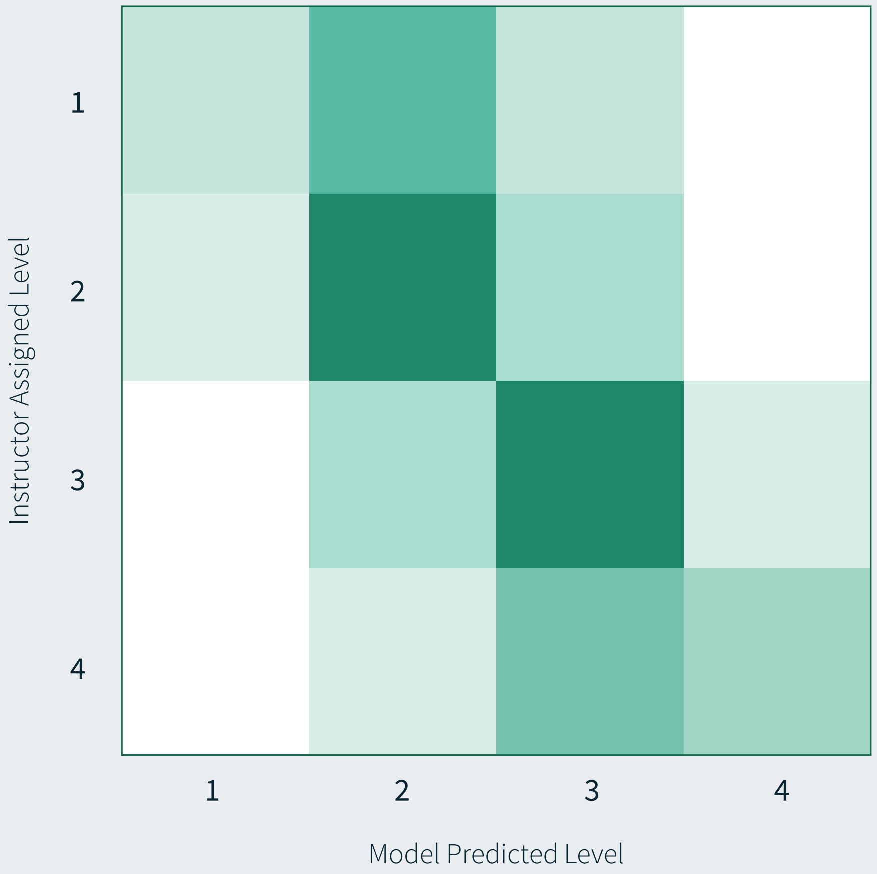


FIGURE 10. Comparison of scores (levels) instructors assigned to students' English compositions with scores predicted using machine learning techniques based on their comments. Higher scores were predicted better than lower scores (darker colors show a higher concentration of data).

Leveraging Machine Learning to Provide Automated Score Recommendations

To what extent is it possible to automatically assign learning objective scores to English compositions based on instructor comments? For this purpose, a machine learning model was developed to predict the assignment of a learning objective score (level) using the content in instructor comments as predictors. Use of several different machine learning techniques (naive Bayes, random forest, support vector machines, multilayer perceptron, and convolutional neural network) demonstrated that the predictability of scores can be improved by about 13% by utilizing convolutional neural networks over naive Bayesian networks. At 62%, the convolutional neural network technique provided the highest prediction accuracy. Higher scores can be predicted better than lower scores (see Figure 10). Although accuracy needs to be improved further for practical use, machine learning techniques show promise. In the future, these techniques may be able to supplement the support instructors offer to their students—for example, provide instructors with prompts for where to re-evaluate or verify an assigned score, or identify compositions with a common problem and recommend feedback to help them provide richer and more personalized feedback in large classes.

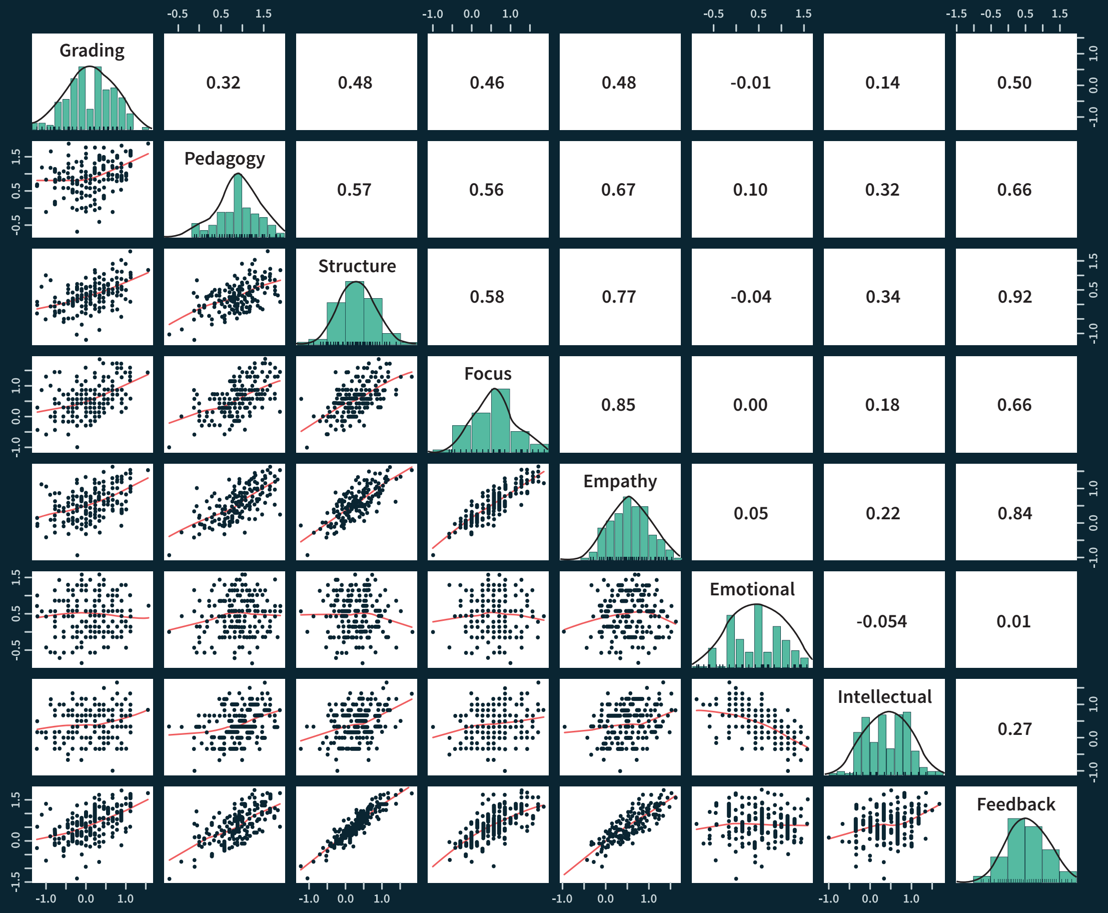
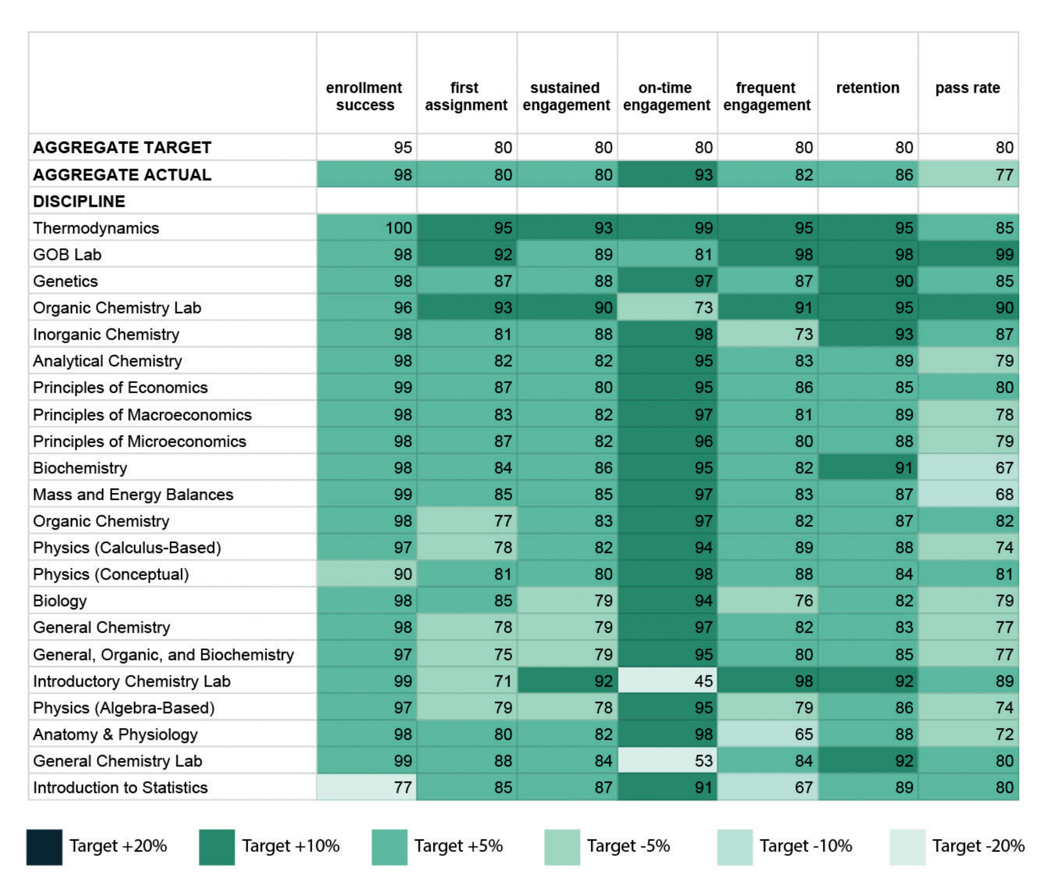


FIGURE 11. Associations of instructor preferences along eight traits. Correlation coefficients between traits are listed in the right half of the grid. "Course structuring" and "timely feedback" are strongly and positively correlated. That is, instructors who value providing timely feedback also prefer structuring their courses in specific ways such that the two are positively associated. In contrast, "emotional drive" and "intellectual drive" are negatively correlated. That is, instructors who value emotional drive tend not to value intellectual drive and vice versa.

Building Educational Products with Empathy: Empirical Analysis of Instructor Personas

Understanding the needs of educators and their value preferences in instruction is essential to the user experience and product development teams at Macmillan Learning in order to develop products that are empathetic and effective. The insights learning analytics can provide extend beyond data generated in online learning platforms, to data from other sources - such as surveys. Extracting insights from responses to well-designed survey questions is non-trivial. Actionable insights from responses to survey questions that attempt to capture the complexities of instructor preferences can only emerge by looking at the data conditionally from multiple angles and utilizing a multitude of approaches from point statistics to associations (see Figure 11) to probe apparent clusterings. The analytics provide insights on the relationships between preferences of different users. In addition, these insights are also used to continuously refine the survey design.



	enrollment success	first assignment	sustained engagement	on-time engagement	frequent engagement	retention	pass rate
AGGREGATE TARGET	95	80	80	80	80	80	80
AGGREGATE ACTUAL	98	80	80	93	82	86	77
DISCIPLINE							
Thermodynamics	100	95	93	99	95	95	85
GOB Lab	98	92	89	81	98	98	99
Genetics	98	87	88	97	87	90	85
Organic Chemistry Lab	96	93	90	73	91	95	90
Inorganic Chemistry	98	81	88	98	73	93	87
Analytical Chemistry	98	82	82	95	83	89	79
Principles of Economics	99	87	80	95	86	85	80
Principles of Macroeconomics	98	83	82	97	81	89	78
Principles of Microeconomics	98	87	82	96	80	88	79
Biochemistry	98	84	86	95	82	91	67
Mass and Energy Balances	99	85	85	97	83	87	68
Organic Chemistry	98	77	83	97	82	87	82
Physics (Calculus-Based)	97	78	82	94	89	88	74
Physics (Conceptual)	90	81	80	98	88	84	81
Biology	98	85	79	94	76	82	79
General Chemistry	98	78	79	97	82	83	77
General, Organic, and Biochemistry	97	75	79	95	80	85	77
Introductory Chemistry Lab	99	71	92	45	98	92	89
Physics (Algebra-Based)	97	79	78	95	79	86	74
Anatomy & Physiology	98	80	82	98	65	88	72
General Chemistry Lab	99	88	84	53	84	92	80
Introduction to Statistics	77	85	87	91	67	89	80

■ Target +20% ■ Target +10% ■ Target +5% ■ Target -5% ■ Target -10% ■ Target -20%

FIGURE 12. Student outcome journeys for a range of disciplines Macmillan Learning provides online learning products for. The metrics are assessed using proxies and provide an overall health check of instructor and student experiences. The heatmap guides opportunities, such as different or additional instructor support, best practices in course design, and product or user experience refinements.

From Instructor and Student Journeys to Better Course Outcomes: Identifying Opportunities for Product Refinement, Support, and Training

In addition to deep analyses of users of specific products, features, and content, analysts at Macmillan Learning have found it valuable to research aggregated metrics of user journeys and key steps related to better course outcomes. The user journeys provide insights into instructors' and students' use of a product in their course throughout a term. For instructors, these analyses reveal how easily they enroll students and create a course, the frequency of online assignments, and course outcomes. For students, these analyses reveal how easily they registered and their level of engagement and performance during the course (see Figure 12). These analyses of user journeys provide insights for product refinement and innovation as well as instructor support and training.

> “The analyses of user journeys provide insights for product refinement and innovation as well as instructor support and training.

The Future of Data Analytics in Education: Reasons for Optimism

Analyzing data from educational contexts is often viewed as a means to summarize observations and test hypotheses. At Macmillan Learning, we take a much broader view of the potential of learning analytics to improve student, instructor, and institutional outcomes. In particular, we have identified a wealth of emerging opportunities for learning analytics to provide empirical insights into foundational questions in education such as learning at an impasse, scaffolding, spaced repetition, multimedia learning principles, and knowledge and problem-solving transfer (Bransford, Brown, & Cocking, 2000; Kahney, 1993; Mayer, 2009; VanLehn et al., 2003). This approach to learning analytics is possible for three key reasons:

> **We have identified a wealth of emerging opportunities for learning analytics to provide empirical insights into foundational questions in education.**

1. the increasing breadth, depth, and timeliness of data from learning platforms,
2. the increasing granularity of data from well-designed learning platforms (e.g., interactions with time stamps), and
3. the ability to factor in contextual details from on-the-ground impact research studies, which provide rich insight into educational processes in their natural settings that are otherwise difficult to obtain—and may even be intrusive—with studies at scale.

Needless to say, these present opportunities and challenges. The increasing availability of a wealth of data has the potential to drown analysts in a sea of information without much meaning. The granularity of data can cause analysts to lose sight of the big picture. Educational impact studies are limited by various course constraints such as course lengths, sizes, and unique contexts that may not be easily repeatable; and control groups may not be possible, limiting causal analyses. However, we contend that all three, when tackled with the best practices encouraged earlier in this paper, have the potential to enable learning analytics to significantly contribute to foundational research in education. It is for these reasons we believe we are entering an age when learning analytics has the potential to become a major contributor to the improvement of education. We are at the cutting edge of an interdisciplinary field including research in teaching and learning, cognitive science, educational impact research, user-centered design, machine learning, and artificial intelligence.

Indeed, a decade and a half ago, in the article, “The frontier of web-based instruction,” Coral Mitchell, Tony DiPetta, and James Kerr wrote (Mitchell, Dipetta, & Kerr, 2001):

“[The] question of how web-based education affects teaching and learning remains largely unanswered, and the terrain of online learning remains largely unmapped.”

Despite much progress, we believe that this observation still holds, and we are on a journey to explore this terrain together with institutions, instructors, students, and other education stakeholders. We recognize that innovations in educational technology provide us with novel opportunities to understand the processes of learning and teaching and, in turn, help us develop educational solutions that meet the needs of instructors and students. Although constraints and limitations apply, as with any endeavor, we believe that learning analytics can leverage data insightfully and empathically—that with sufficient and up-to-date data, we can support students and instructors to achieve greater outcomes.

> “We believe we are entering an age when learning analytics has the potential to become a major contributor to the improvement of education.

We look forward to the journey ahead.

Analyzing Data Responsibly: Safeguarding User Privacy

At Macmillan Learning, data analytics go hand-in-hand with the utmost respect we have for learners, instructors, and administrators. To protect personally identifiable data, we apply administrative, technical, and physical security measures. We also comply with all relevant privacy laws and our agreements with educational institutions. See our Privacy Notice at **https://store.macmillanlearning.com/us/privacy-notice** for more information on how we handle data.

Safeguarding user privacy is paramount to us while conducting data analytics at Macmillan Learning. Our data analytics, including the examples in this paper, are conducted without associating the analytics data with the respective users' names or other personally identifiable information.

Data analytics at Macmillan Learning is conducted for the sole purpose of improving our products and services to enhance the educational outcomes of students and instructors. We regularly monitor our ethical use of data and take continual steps in implementing best practices to safeguard user privacy in consultation with legal and privacy experts.

REFERENCES

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

Brooks, R. (2017). The seven deadly sins of AI predictions. *MIT Technology Review*, October.

Freeman, S. et al (2014). Active learning increases student performance in science, engineering, and mathematics. *PNAS*, May.

Gould, S. (1991). “The median isn’t the message” from *Bully for Brontosaurus: Reflections in Natural History*. W.W. Norton & Company.

Gyurko, J., MacCormack, P., Bless, M.M., & Jodl, J. (2016). *Why colleges and universities need to invest in quality teaching more than ever*. Washington, DC: American Council on Education, & Association of College and University Educators.

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hayes, B. (2014). Doing data science. *Notices of the AMS*, 61(9), 1068-1071.

Kahney, H. (1993). *Problem solving: current issues*. Open University Press.

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist*, 63, 760-769.

Mayer, R. E. (2009). *Multimedia learning*. Cambridge University Press.

McKinsey Global Institute (2016). *The age of analytics: competing in a data-driven world*.

McWilliams, K. et al. (2017). Unpacking the black box of efficacy, Macmillan Learning White Paper.

Mitchell, C., Dipetta, T. & Kerr, J. (2001). The frontier of web-based instruction. *Education and Information Technologies*, 6, 105-121.

Seaman, J., Seaman, J. (2017). *Opening the Textbook: Educational resources in U.S. Higher Education 2017*. Babson Survey Research Group.

Silver, D. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529, 484-489.

VanLehn, K. et al. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21, 209-249.

About the Authors

Dr. Rasil Warnakulasooriya
Vice President, Learning Analytics

Rasil studied physics at the University of Colombo, Sri Lanka, at Rice University, and at The Ohio State University. He spent his postdoctoral days at the Massachusetts Institute of Technology researching online learning with Professor David Pritchard. He has enjoyed building and leading analytics divisions in several companies, taking novel approaches to data analytics including predicting learners at risk in science using a fractals-based approach (for which he was awarded a patent), to researching the micro impact of individual learning activities, to identifying empirical differences of English- language learners around the world. Throughout, Rasil is driven by a passion for extracting meaningful insights into the subtleties of learning from complex and messy data.

Dr. Adam Black
Chief Strategy & Learning Officer

Adam is a recognized pioneer in improving learner outcomes. From identifying promising applications of learning science, through directing the development of market-leading digital products (used by more than 25 million learners, at the time of writing), to spearheading novel approaches to assessing and improving impact, Adam has been dedicated for 25 years to helping instructors and institutions to improve student success. Adam holds a BSc in Physics from the University of Edinburgh and a PhD in Astrophysics from the University of Cambridge, has a patent for predicting learners at risk, and has won national and global awards for digital product innovation.

About Macmillan Learning

Macmillan Learning improves lives through learning. Our legacy of excellence in education informs our approach to using user-centered design, learning science, and impact research to develop world-class content and pioneering products that are empathetic, highly effective, and drive improved outcomes. Through deep partnership with the world's best researchers, educators, administrators, and developers, we facilitate teaching and learning opportunities that spark student engagement and lift course results. We provide educators with tailored solutions designed to inspire curiosity and measure progress. Our commitment to teaching and discovery upholds our mission to improve lives through learning. To learn more, please visit **www.macmillanlearning.com** or see us on Facebook, Twitter, LinkedIN or join our Macmillan Community.

About the Learning Science and Insights Team

As the Learning Insights company, we are passionate and scientific about helping students, instructors, and institutions to achieve their full potential. We use a unique combination of user-centered design, research from the learning sciences, and empirical insights from extensive data analytics and Impact Research. To learn more about the Learning Science and Insights team please visit **www.macmillanlearning.com/learningscience**

macmillan learning