UNPACKING THE BLACK BOX OF EFFICACY

A framework for evaluating the effectiveness and researching the impact of digital learning tools.

Kara McWilliams // Adam Black // Jeff Bergin // Rasil Warnakulasooriya

macmillan learning

This White Paper cuts straight to the debate we should all be focused on – to ensure that future learning products are even more powerful, we need research to understand not just whether, but why a learning experience has efficacy."

- Dr. Chris Dede

Timothy E. Wirth Professor in Learning Technologies, Technology Innovation, and Education Program at the Harvard Graduate School of Education.

This White Paper offers clear and critical explanations of RCTs and other educational research methods Macmillan employs as well as a reasonable and practical proposal for how the company intends to combine their respective strengths and weaknesses."

- Michael Feldstein

Partner at MindWires Consulting, Co-Publisher of e_literate, and Co-Producer of e-Literate TV.

This White Paper proposes what should be adopted as best practice for studying the effectiveness of digital learning tools on teacher pedagogy and student engagement and learning – it proposes a practical and rigorous blend of multiple methods, including experimental and quasi-experimental designs, case studies, and data analytics."

- Dr. Suzanne Lane

Department Chair Research Methodology, University of Pittsburgh.

Contents

Foreword

Digital learning tools have significant potential to contribute to improving outcomes in higher education. However, great learner outcomes are the result of an educational ecosystem and a learning product is just one component. We believe that any effort to measure the effectiveness of a tool must, therefore, take into account the context.

Impact research has traditionally focused on tightly controlled trials. However, these often take years to execute, don't accommodate the important and real variability of learners and their educational settings, and provide results too late or too slowly to be useful to instructors and education institutions or contribute to the iterative and ongoing improvement of a learning product.

In this paper, we propose a different approach: a framework for evaluating the effectiveness of digital learning tools that begins during development and continues once a product is in use. The framework starts by building a clear understanding of the variety of ways users choose to use a learning product, and progresses through increasingly rigorous studies, repeated across different educational environments and use cases. This approach provides a continuously growing body of evidence that, we believe, provides more relevant and reliable insights into how a product will be effective and under what circumstances.

Measuring effectiveness is fundamentally difficult. At Macmillan Learning, we do not claim to have all of the answers. But we do know that instructors want to make more informed decisions on what products to use and how to use them most effectively with their students and in their particular educational setting. We therefore endeavor to provide instructors with evidence that is practical, relevant, transparent, and reliable. We take that responsibility very seriously, and are excited to contribute. We thank the expert advisors and instructors working with us. We look forward to continued and on-going engagement with the educational community and hope that this White Paper makes an interesting and useful contribution.

ACKNOWLEDGEMENTS

This work benefited immeasurably from the thoughtful guidance and careful critique of the **Macmillan Learning Impact Research Advisory Council (IRAC)**. The Council comprises a diverse panel of external experts in designing and measuring the impact of educational technology, methods for measuring effectiveness, modeling and evaluating learning performance, standards for measurement in education, and respecting current and evolving data privacy standards and laws. The Council members are: Dr. Chris Dede, Timothy E. Wirth Professor in Learning Technologies, Technology Innovation, and Education Program at the Harvard Graduate School of Education; Michael Feldstein, Partner at Mindwires Consulting; Dr. Sara Finney; Professor, Department of Graduate Psychology, and Associate Director in the Center for Assessment and Research Studies, James Madison University; Dr. Suzanne Lane, Department Chair Research Methodology, University of Pittsburgh; Dr. Thanos Patelis, Research Scholar at Fordham University and Principal Scientist at Human Resources Research Organization; and Dr. Elana Zeide, Yale Law School Visiting Fellow, Information Society Project; Princeton University, Associate Research Scholar, Center for Information Technology Policy.

To each of you, our sincerest thanks.

We want to acknowledge the contributions from our colleagues across Macmillan Learning whose experiences, expertise, and passion to help students, instructors, and institutions to succeed helped shape this framework. In particular, we want to thank the enthusiastic support of Ken Michaels, our CEO, whose passion for helping more students succeed led him to believe that measuring effectiveness should become fundamental to how we do business. We also want to thank the Communications and Design team for their help producing this paper.

Finally, we are indebted to the education institutions, instructors, and students who partner with us on every aspect of product design, development, testing, and impact research. Their generous contributions, honest feedback, keen insights, and critical perspectives help us to continually evolve and re-evaluate how we can best support instruction and learning in higher education.

Introduction

Instructors are flooded with choices as new digital learning tools enter the higher-education market. To avoid false starts, frustration, and missed opportunities for a class or group of students, they need to know which really contribute to learning, to what degree, for whom, and in what context.

With such high stakes, we encourage institutions, instructors, and students to demand more transparent, reliable, and relevant evidence so they can make the best informed decisions about what learning products to use, why, and how.

Unfortunately, many of the currently available digital learning tools lack evidence of their effectiveness. Dr. Robert Pianta described the current use and evaluation of digital learning tools as "at best, we are throwing spaghetti against the wall and seeing if it sticks, except that we don't even know what it means to stick" (**EdTech Efficacy Symposium, 2017**). Where supporting research does exist, it often relies on traditional

> With such high stakes, we encourage institutions, instructors, and students to demand more transparent, reliable, and relevant evidence so they can make the best informed decisions about what learning products to use, why, and how.

methods that evaluate use and outcomes in a unique setting, or broadly comparing outcomes between users and non-users in rigorous longitudinal trials, but ignoring differences in contexts in which they're used. Isolated statements of efficacy may not be the most meaningful way to help decision makers. However, innovative approaches to effectiveness and impact research, and a reconsideration of the "gold standard" of research, can open up insights for instructors and learners that are practical, actionable, and timely.

We believe that instructors and learners will find a comprehensive narrative illustrating whether a product works, for whom, when, and in what contexts is more useful than traditional control trials that try to isolate the impact of a tool on a specific learner outcome, for a specific use case, and in a unique educational context. An evolving portfolio of evidence beginning in development and continuing as a product matures, offers the opportunity for interconnected findings from educational research, learning science, data analytics, and local, contextual studies to provide more relevant, reliable, and actionable insights to institutions, instructors, and learners.

Building an evolving portfolio of evidence of effectiveness and impact is complicated. Doing it well requires an open mind, understanding about the limitations of conducting research in educational ecosystems, on-going partnership with leading experts in the field, and on-going engagement with instructors and students using the products to refine the approach based on their feedback.

At Macmillan Learning, we are committed to improving learner success and supporting faculty. So, as we evolve this dynamic approach to studying the effectiveness and impact of our digital learning tools, we aim to be transparent about our methods and results so users can understand and be confident in the insights we are sharing. We acknowledge that the framework has limitations and we look forward to ongoing collaboration with the faculty and learners who use our tools and expert academics to continually refine this approach. As a first step, we have developed this White Paper to share our approach.

> “…a comprehensive narrative illustrating whether a product works, for whom, when, and in what contexts is more useful than traditional control trials.”

To set context, this paper begins by outlining approaches that are traditionally used to measure the impact of digital learning tools, and the challenges and limitations of each. We then discuss the research and evaluation taxonomy that our framework is built upon, and we present the framework and its limitations. The paper concludes with our ongoing efforts to seek guidance and continuously refine these methods.

Current Approaches to Measuring the Efficacy of Digital Learning Tools

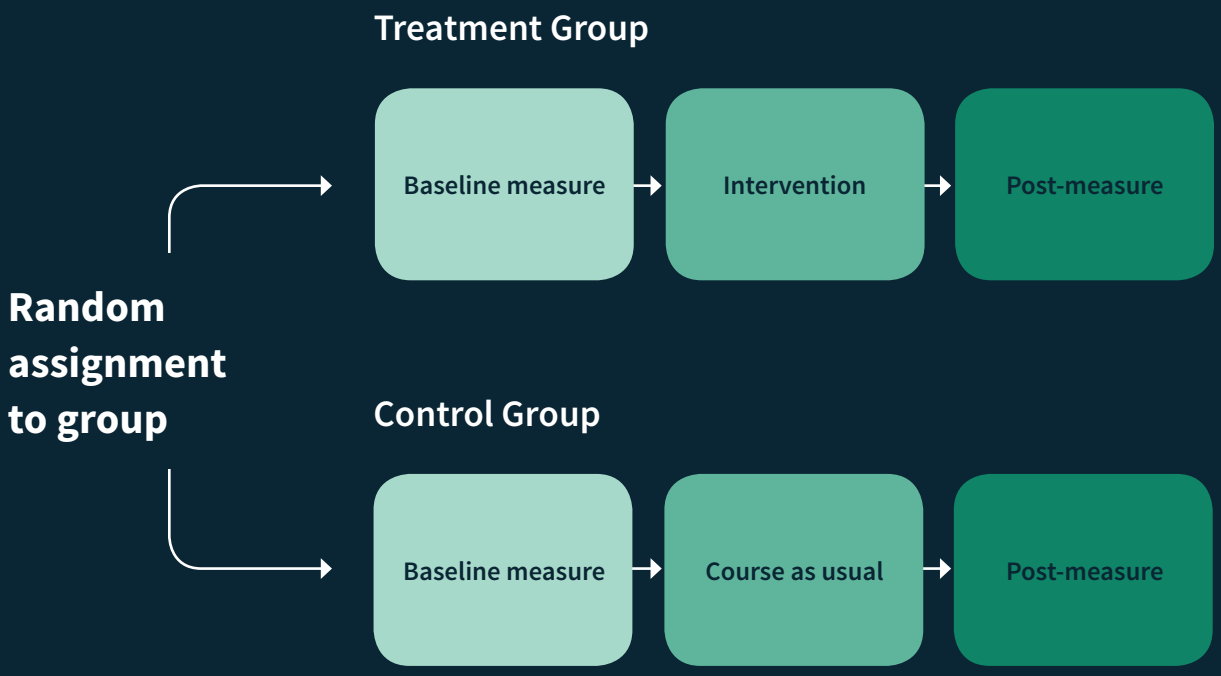
Although many of the digital education products currently available lack evidence of effectiveness, researchers and evaluators have been working to understand how to best measure their efficacy. The gravity of these efforts is made obvious through publications like the **United States Department of Education's Expanding Evidence Approaches for Learning in a Digital World (2013)** and initiatives like the **2017 Edtech Efficacy Research Academic Symposium** where nearly 200 stakeholders convened for two days to review the year-long efforts of working groups whose goal is to progress efforts in efficacy measurement. At that meeting, a consensus emerged that more innovative methods are needed. Despite this, researchers continue to rely heavily on traditional research and evaluation techniques to try to understand 'what works'.

Two research methods are typically implemented to measure the efficacy of digital learning tools: experimental designs and case studies. The growth of data capture in online digital learning products has also driven a third: exploratory data analytics. Each method has utility. But, conducted in isolation, none can comprehensively evaluate a product's effectiveness or measure its impact on instructor and learner outcomes in a way that provides users with relevant, reliable, timely, and actionable insights.

THE EXPERIMENTAL DESIGN

The experimental design has long been considered a **"gold standard"** for research in medicine and psychology because study results are assumed to isolate an intervention as the cause of an outcome. In these fields, an experimental design is represented as a study conducted in highly controlled laboratory settings with participants being randomly assigned into either the treatment group (receiving the studied treatment), or the control group (receiving a placebo). The tight controls and randomization garner the design name, randomized control trial (RCT).

When used in education, the same gold standard is assumed: a sample of students representative of the population of interest is identified and within that sample students are randomly assigned to either the treatment group (receiving an intervention), or the control group (receiving the course experience as usual).



Current Approaches to Measuring the Efficacy of Digital Learning Tools

Because randomization is assumed to account for pre-existing differences between groups, and controlled conditions account for their experience during the experiment, null hypothesis significance testing (NHST) can isolate the impact of a tool or intervention. Using NHST, a researcher can explore whether the outcome of interest is different enough between groups to conclude that the intervention caused the difference. It is assumed that a significant difference in the outcome suggests that the intervention worked, and that the same results can be reasonably expected in similar settings - without the need to replicate the study.

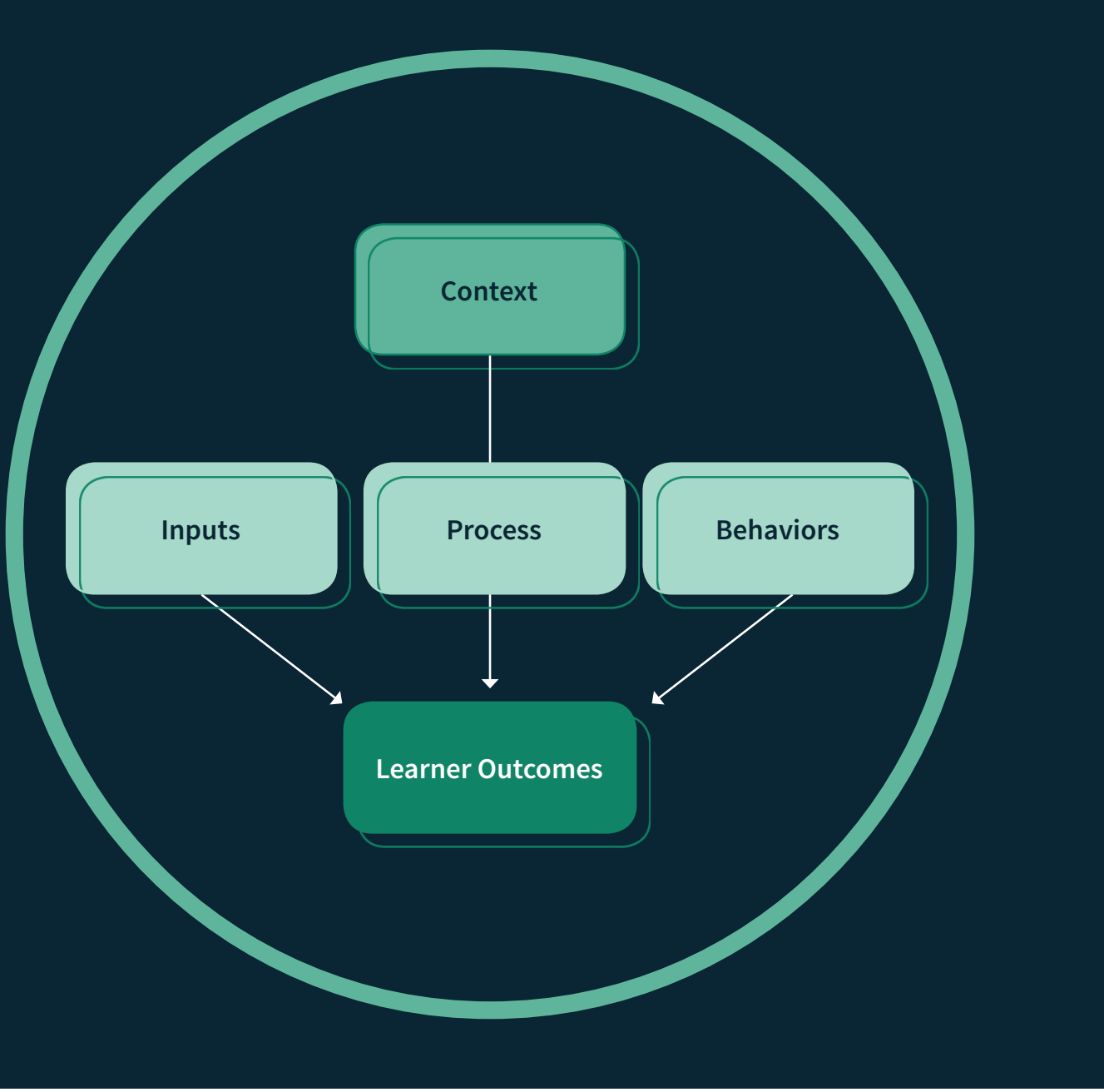
The complex context in which educational RCTs are conducted, however, challenges many of the assumptions upon which they are based. Well-designed RCTs assume that they are executed in a highly controlled laboratory setting, but educational research is typically conducted in a highly contextualized and largely uncontrolled ecosystem. Even randomization cannot account for differences in learning environments or how instructors choose to use the learning tool.

> "The complex context in which educational RCTs are conducted… challenges many of the assumptions upon which these methods are based."

Institution, course, faculty, and learner characteristics all contribute to how a tool will impact learner outcomes. Failing to consider those contexts and implementations (that is, how a product is used) when measuring effectiveness risks results that are biased and insights that may not be relevant to a large proportion of users. For example, results from an RCT conducted at a large, four-year, private institution, where the product is used as a supplemental learning tool may not be directly relevant to faculty who teach at small, two-year, community colleges, where the product is used as the primary learning tool. Similarly, RCTs eliminate the opportunity to support faculty implementation decisions because results indicate that a product impacts an outcome but not why. A statistically significant change in a measured outcome does little to help faculty understand what use cases will lead to those results and how to refine their course design and delivery.

RCTs also assume that the intervention being tested remains consistent for the duration of the study. But, controlled trials take a long time to execute and analyze - sometimes years - and many online digital learning tools are continually being refined and evolved. If users are engaging with a product as it evolves, it is very difficult to isolate the impact of the product on the outcome being studied. Even if researchers can control for different contributors to and drivers of outcomes, by the time the study findings become available the digital tool may have evolved so much that the results are no longer relevant.

RCTs are an effective method for controlling for factors that may contribute to an outcome being measured, but the complex ecosystem in which learning takes place makes it nearly impossible to isolate the impact of a tool or intervention on learner outcomes. And because results of RCTs take so long, when conducted in isolation, they do little to support institutional and faculty adoption and use decisions.



THE CASE STUDY

The case study is an approach to measuring effectiveness where an intervention is examined in its real-world setting with little or no modifications to naturally occurring use. Case studies are not bound to any particular data-collection methods. The type of data that are collected are driven by the intervention being studied and may include: product data, classroom observations, instructor and learner interviews, focus groups, and other course artifacts.

The case study focuses on a holistic description and explanation of the intervention, with the researcher often making some judgment about the effectiveness of the intervention being studied. In the case of the implementation of a digital learning tool, a researcher would consider the context of the educational environment in which it was used, any other inputs into the course (such as other tools an instructor is using), how the instructor uses the tool, and the behaviors of the learners using the tool (e.g. classroom and course engagement). All of these data are synthesized and the researcher makes a judgement about the impact of the tool on the outcome or outcomes of interest in the specific, local context.

Current Approaches to Measuring the Efficacy of Digital Learning Tools

Typically, a case is chosen as an instance of some hypothesis or concern. For example, usage data may suggest that learners are highly engaged using a tool, or realizing outcomes that are higher than expected. To understand what is influencing these results, researchers need insights from a local examination of one or more courses and classes.

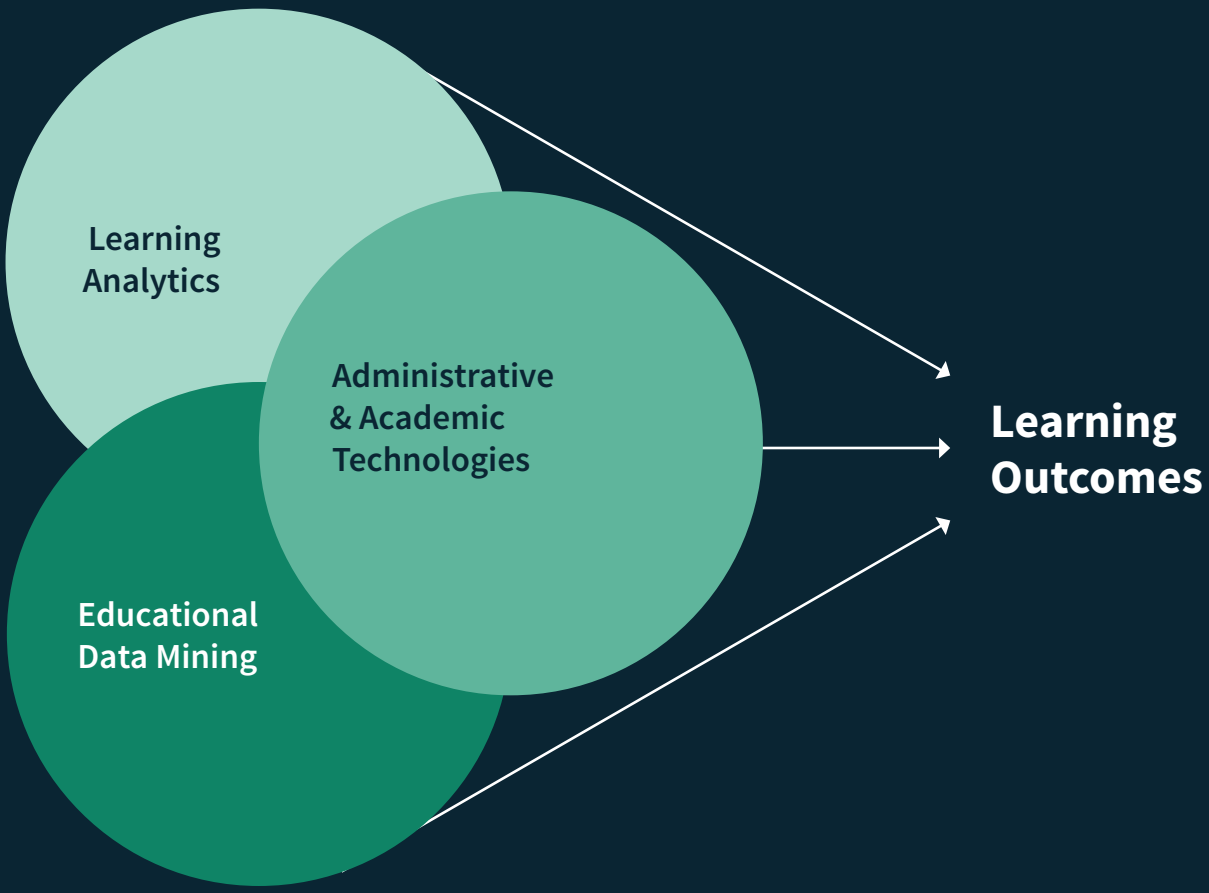
The case study is a more sensitive form of measurement - one that generates richer results, offers insights that expand experiences, and furthers the knowledge base about a particular phenomenon. This method is particularly useful for evaluating an educational initiative and improving the use of a tool or design of a program.

> "The case study… is particularly useful for evaluating… and improving use of a tool or design of a program."

The case study addresses the risks of ignoring context, environment and implementation because rich data are captured on each. Nonetheless, the case study does have critical limitations including issues of reliability, validity, and generalizability. The conclusions drawn from the data collected in a case study are subjective and depend on the sensitivity, experience, and even integrity of the researcher. Qualitative results may be oversimplified or exaggerated due to existing biases of the researcher. Controls can be put in place to increase validity and reliability - such as having a second researcher collect data and make judgements - but these are costly and time intensive.

Results from a case study also lack representativeness. Similar to results from experimental designs, how and why something is working in a particular case cannot usually be generalized to a broader population because the contexts will likely differ. In order to increase generalizability beyond a specific case, a stratified sample of the population of interest should be developed, cases from each strata sampled and studied, and a meta-analysis of the results conducted.

Individual case studies are valuable for illustrating to instructors how faculty with similar contexts and students have used a tool and what outcomes they achieved. However, conducted in isolation, a case study doesn't evaluate the overall effectiveness or efficacy of a digital learning tool.



EXPLORATORY DATA ANALYTICS

In this emerging methodology for measuring the impact of digital learning tools, data captured by the tools are used to explore usage and behavioral trends and how they relate to learner outcomes. A combination of embedded administrative and academic technologies and digital learning tools inside and outside the classroom provide the potential for a comprehensive view of if, when, and how a learner is engaging with a tool and their behaviors and performance.

Methods used in exploratory learning analytics are very diverse. In general, data captured as a learner engages with a tool - for example, when and how often they use the product, the time they spend on a particular activity, and item- and test-level assessment scores - are matched with student information systems or other sources of student data.

Analyses may be summative - examining how usage and behaviors through the course are related to final course outcomes, and how the product design may be refined. They can also be dynamic - providing instructors with real-time insights into how students are progressing and possible interventions, or exploring a/b testing of content or product features.

Especially for products used at scale, exploratory analytics can provide powerful insights at the aggregate level and also by segmenting users. However, like the experimental design and the case study, analytics have limitations when conducted in isolation. The fundamental limitation of exploratory analytics to measure effectiveness is that the findings are exploratory, so the resulting trends are suggestive rather than causal. If there are a sufficient number of users, analytics can be used to segment users. However, that segmentation is limited to a large extent by what data is or is not captured by the digital tool (or tools used alongside it). This means that important differences between local contexts and educational environments may not be apparent in the data and thus can't be allowed for.

A COMMON CHALLENGE

These methods are diverse, and have unique benefits and limitations. However, there is an emerging challenge that they all share. Digital learning tools and the introduction of data analytics has expanded opportunities to understand much more about students - who they are, how they learn, and their behaviors while learning. These data can be highly meaningful for faculty, administrators, learners, and researchers. However, wider access to richer learner data also raises privacy risks for the study participants.

Mitigating potential risks to study participants is not a new consideration. But as more comprehensive data are housed electronically, security vulnerabilities become greater. Historically these vulnerabilities have been modest, so laws protecting learner data in the United States have been vague - typically requiring only that reasonable measures are taken to protect student privacy. As more comprehensive data are housed electronically on learners, the nation is trending toward more stringent requirements. However, a lack of nationally agreed upon standards for student data privacy result in inconsistent practice, leading institutions to understandably err on over-protecting student data at the institution level, and this limits the availability of student records to be used for research and improvement purposes.

> “Effectiveness and impact results should answer not only if a product impacts outcomes, but how and why.”

To understand the relationship between product usage and learner outcomes, a researcher typically requires at least some student information to be collected at the institutional level. Obtaining these data requires approval by the institution’s Institutional Review Board (IRB), or a third-party IRB, and sometimes both. As part of these approvals researchers are required to demonstrate that they have the qualifications and credibility to protect student participants, an infrastructure with appropriate data security, and clear, comprehensive standards for handling data. The national trend toward stricter standards around student data privacy and security have led to IRB approval processes that are more complex and take more time and resources to obtain.

TO SUMMARIZE

Researchers *are* working toward understanding the impact that digital learning tools have on instructor and learner outcomes using a range of methods, including experimental designs, case studies, and exploratory analytics. But, when conducted in isolation, these methods typically do not provide instructors and learners with insights that are relevant, reliable, timely, or actionable for their local context.

Effectiveness and impact results should answer not only *if* a product impacts outcomes but *how* and *why*. The results should be communicated in a way that can be used by instructors and other users of the product and timely enough that they can inform adoption and usage decisions. Also, it is critical to put in place processes and systems to guarantee the privacy and security of faculty and student data.

With these considerations in mind, at Macmillan Learning we believe that the effectiveness of a digital learning tool cannot reliably and usefully be defined by a standalone statement of efficacy. Instead, a holistic approach to understanding product use cases and learners, instructors, and their environments is a more meaningful way to provide insights to users to help them to achieve better outcomes. Our goal is therefore to identify appropriate outcomes for the products being examined and build an evolving portfolio of evidence examining if, how, and why a product positively influences those outcomes in a variety of contexts. We propose that such a portfolio should comprise local, user-centered, context-specific data, and national aggregated or segmented data. We also recommend that research and evaluation methods used to examine those data should be appropriately rigorous for the product's stage in a life cycle, but always adhere to measurement standards. We believe that this evolving portfolio of evidence will provide faster and more reliable, relevant, and actionable insights to users and those refining the design of a product; identify and refine implementation models that enable users to achieve the best outcomes in different educational settings; and provide transparent and reliable evidence to support claims about effectiveness appropriate to a product's stage of evolution.

> "…research and evaluation methods… should be appropriately rigorous for the product's stage in a life cycle, but always adhere to measurement standards."

To achieve this we propose a framework for evaluating the effectiveness and researching the impact of digital learning tools. Within this framework, evaluation of effectiveness begins while a product is being developed and continues as it enters into use. Impact research is conducted once a product has matured in use and after the various naturally occurring use cases have been identified and documented. Evidence of a product's effectiveness and impact rests on evaluations and research studies being repeated across varied educational environments and use cases.

Development of this framework began by establishing a research and evaluation lifecycle. We then identified research study designs that would provide the most useful insights at each stage of a product lifecycle. Finally, we organized these studies so that, over time, they build a connected and comprehensive body of evidence for a product.

This framework isn't without limitations, nor is it fixed. We will continue to evolve it working closely with instructors and students using our learning tools, and seeking on-going feedback from expert academic advisors (including specialists in evolving data security legislation). We will also publish updates to this White Paper as the framework evolves to provide transparency into our methods, insights, and claims about our products.

Product Research, Design, and Evaluation Lifecycle

From discovery, where we co-design digital learning solutions with learners and faculty, through delivery where we seek to separate the product's impact from other outcome drivers, the product research, design, and evaluation lifecycle is the foundation of our effectiveness and impact framework.

DISCOVERY

CO-DESIGN & LEARNING RESEARCH

Research performed

Co-designing with users

User and outcome learning research

Questions explored

What are the real-life journeys of faculty and students?

What educational research are we using as the foundation of the product design?

Insights gleaned

Guide product development to create products that are likely to positively influence learner outcomes

Justifiable claims

"Research indicates that pedagogical model A will positively influence learner outcomes X and Y"

Time Required

~3-6 months

DISCOVERY - CO-DESIGN & LEARNING RESEARCH

Product discovery begins with empathy research to deeply understand student and faculty needs, contexts, goals, and challenges. From there, we collaborate on ideation – brainstorming ideas and solutions to help them to solve their biggest problems and achieve their ambitions in the most practical and efficient ways. Then we critically assess and synthesize research in the learning sciences (that is, education research and cognitive science) to guide the solution design. These "design principles" focus on motivation, cognition, and pedagogy. We enrich this with empirical insights from novel and extensive data mining research collaborations with faculty and institutions. This research often reveals nuanced differences in student behaviors and needs.

> "Product discovery begins with empathy research to deeply understand student and faculty needs, contexts, goals, and challenges."

During discovery, intended student, instructor, and institution outcomes are identified for a product. The development of measureable outcomes begins with an outcomes framework that underpins all Macmillan Learning products. Outcomes at the discipline and product level are then developed in partnership with product teams and the instructors and students who co-create with us. Researchers then establish leading indicators and metrics by which to measure them. These help instructors and other stakeholders to know that the product is influencing important outcomes.

LEARNING DESIGN & DEVELOPMENT

Research performed

Usability tests

Pilot tests

Questions explored

What is the user experience?

How might it be improved?

What product effects can we see in a small, controlled setting?

Does a specific feature influence outcomes?

How can design or use be optimized?

Insights gleaned

Recommendations to improve the user experience for all users or specific use cases

Information to support instructor or institution adoption decisions

Justifiable claims

“Feature A distracted students from the content”

“Feature B encouraged student engagement and progression through the course”

“Testing results suggest use of the product is related to learning performance in X class”

Time Required

1 semester

Product Research, Design, and Evaluation Lifecycle

LEARNING DESIGN AND DEVELOPMENT

Foundational evidence gathered during discovery is used as we ideate new, novel solutions and prototype and test them with students, instructors, and other faculty. This iterative refinement results in solutions that are research-based and user-centered so they are impactful, intuitive, and highly usable. We work with a wide variety of users to explore how they engage with components of a product - at one of the Macmillan learning labs, in a remote testing space, on campus to participate in or observe one of their courses. Testing during development helps us to understand the user experience for students and instructors and how it might be improved.

Later in development, we invite users from various segments and representing our target personas to use the product for a complete semester. During these beta tests we document implementation and collect systematic data on the user experience and student and instructor outcomes. We learn what product effects can be observed in small settings with deeply understood contexts and evaluate whether use of specific product features is related to outcomes. From these tests, students and instructors help us to identify how to improve the user experience and overall product design for all users and for specific use cases.

OUTCOME LADDER

Outcome for learner

Advancement
Learner **advances** to next course of study

Learning Performance
Learner **achieves** competence, skills, and confidence

Engagement
Learner **progresses** through material as intended

Access
Learner **onboards** smoothly with supports they need

ADOPTION & OPTIMIZATION

Research performed

Implementation studies

Rapid-cycle evaluations

—

Questions explored

What are typical use cases? How do they vary?

Do learner outcomes vary by use case?

How does the product influence outcomes?

What specific features or capabilities appear to be related to learner outcomes?

—

Insights gleaned

Which use cases lead to best results to guide instructor training and support

Further recommendations for optimizing product design

Additional insights to further support adoption and usage decisions

—

Justifiable claims

"Use of product was related to better learner outcomes when used [this way]"

"Used [this way, e.g. assignment before class] in [this context] was related to better outcomes."

"This feature [e.g. pre-lecture] was related to more active learning in [this context]"

—

Time required

~1 semester per use case

~8 weeks per rapid-cycle evaluation study

ADOPTION AND OPTIMIZATION

As instructors choose to use the digital product in their courses, their natural, local usage decisions are meticulously documented. Through carefully implementation studies, we explore variations in how students, instructors, and institutions use a product and any differences in outcomes they achieve as a result. This is used to further refine the product design and guide support and training to help students and instructors to achieve their best outcomes.

At this stage we also partner with instructors to conduct rapid-cycle evaluations of specific product features. We explore how these contribute to instructor and learner success, identify opportunities for further feature optimization, and identify how to help future students and instructors to understand the utility of various features within a product.

> Through carefully designed studies, we explore variations in how students, instructors, and institutions use a product and any differences in outcomes they achieve as a result."

IMPACT RESEARCH & EVOLUTION

Research performed

Impact studies

Meta-analyses

Questions explored

Does the product produce the desired outcomes when used [this way]?

Are results similar for different subpopulations?

How do results compare with/out use?

Insights gleaned

Rigorous evidence demonstrating outcomes achieved in a variety of educational settings

Rigorous evidence isolating the impact of the product on learner outcomes

Justifiable claims

"Students who used Product X achieved better outcomes than like peers who did not"

"Product X improved retention and completion [across two- and four-year institutions]"

Time Required

>1 year

IMPACT RESEARCH AND EVOLUTION

Once a product has been used in live courses for at least a year, has been refined and optimized using results from previous studies, and instructor use cases have been been systematically categorized, we partner with institutions to conduct impact evaluations. For these, we identify the outcomes that are most important at those institutions and design rigorous studies that help to separate the influence of the product from other outcome drivers. We then repeat the study at other institutions with similar use cases. Finally, we perform meta-analyses across the results from studies at different institutions and within different use cases.

Results from impact studies and meta-analyses provide evidence for if, how, and why the product influences the desired outcomes in a variety of contexts and settings and how those may vary among subpopulations. They crucially rest upon insights and context provided from previous rapid-cycle evaluations and implementation studies, including representative users and use cases. In this way, institutions and instructors can make informed decisions and have confidence that the product and chosen use case is most impactful for their students.

> "Results from impact studies… provide evidence for if, how, and why the product influences the desired outcomes… and rest upon insights and context provided from previous… studies"

The Macmillan Learning Product Research, Design, and Evaluation Lifecycle

Discovery - Co-design & Learning Research	Learning Design & Development	Adoption & Optimization	Impact Research & Evolution
RESEARCH PERFORMED			
• Co-designing with users • User and outcome learning research	• Usability tests • Pilot tests	• Implementation studies • Rapid-cycle evaluations	• Impact studies • Meta-analyses
QUESTIONS EXPLORED			
• What are the real-life journeys of faculty and students? • What educational research are we using as the foundation of the product design?	• What is the user experience? • How might it be improved? • What product impacts can we see in a small, controlled setting? • Does a specific feature influence outcomes? • How can design or use be optimized?	• What are typical use cases? How do they vary? • Do learner outcomes vary by use case? • How does the product influence outcomes? • What specific features or capabilities appear to be related to learner outcomes?	• Does the product produce the desired outcomes when used [this way]? • Are results similar for different subpopulations? • How do results compare with/out use?
INSIGHTS GLEANED			
• Guide product development to create products that are likely to positively influence learner outcomes	• Recommendations to improve the user experience for all users or specific use cases * Information to support instructor or institution adoption decisions	• Which use cases lead to best results to guide instructor training and support • Further recommendations for optimizing product design • Additional insights to further support adoption and usage decisions	• Rigorous evidence demonstrating outcomes achieved in a variety of educational settings • Rigorous evidence isolating the impact of the product on learner outcomes
JUSTIFIABLE CLAIMS			
• "Research indicates that pedagogical model A will positively influence learner outcomes X and Y"	• "Feature A distracted students from the content" • "Feature B encouraged student engagement and progression through the course" • "Testing results suggest use of the product is related to learning performance in X class"	• "Use of product was related to better learner outcomes when used [this way]" • "Used [this way, e.g. assignment before class] in [this context] was related to better outcomes." • "This feature [e.g. pre-lecture] was related to more active learning in [this context]"	• "Students who used Product X achieved better outcomes than like peers who did not" • "Product X improved retention and completion [across two- and four-year institutions]"
TIME REQUIRED			
• ~3-6 months	• 1 semester	• ~1 semester per use case • ~8 weeks per optimization study	• >1 year

A Framework for Evaluating Effectiveness and Measuring Impact

With the research, design, and evaluation lifecycle developed, we collaborated with expert academics to refine the stages and research performed at each. We then overlaid the framework for evaluating effectiveness and researching impact - which outlines the timeline and types of studies - onto the lifecycle.

The evidence resulting from following this framework documents: the educational research underpinning the product design; how the product was refined and optimized based on instructor and learner performance, behaviors, and feedback; how use of the product is related to key outcomes, in a range of contexts; and that research and evaluation methods were appropriately rigorous and met standards of measurement. We hope that this portfolio of evidence provides instructors and institutions with more relevant, appropriate, reliable, transparent, and timely insights to help their decision making about if, how, and why a product is the best solution for them and their students.

The following graphic is a visual representation of the framework followed by a brief overview of study designs and data collection instruments.

Building a Portfolio of Connected Evidence

APPROVAL BY INSTITUTIONAL REVIEW BOARDS

It is essential that institutions and instructors are confident that the studies being conducted are ethical and pose no material risk to participants. Prior to engaging with instructors, all studies are assessed by a third-party accredited review board. A researcher's credentials, study designs, consent forms, instruments, incentives, and data-handling process and procedures are reviewed and any required modifications are made. Once an instructor agrees to partner on a study, we engage with their institution's IRBs and make any further modifications to the study design or processes as necessary.

STUDY DESIGNS

Beta Testing

Data collected during beta testing helps to optimize the product being studied and provide evidence about the product's effectiveness before being launched. Quantitative and qualitative data are collected from instructors and consenting students at various time points over the course of a semester. Descriptive and correlational analyses are then conducted, including: how the tool is naturally used in the course; instructor and student perceptions of the tool; a comparison of student and instructor perceptions of and comfort with technology before and after using the tool; a comparison of student motivation before and after using the tool; relationships between use of tool and learner outcomes; relationships between the use case and learner outcomes; and relationships between instructor perceptions and learner outcomes. Results from the beta tests are shared with the participating instructors and institutions and Macmillan Learning teams responsible for refining and optimizing the product. Case studies and research reports are developed in collaboration with participating instructors so they have a systematic understanding of the effectiveness of the tool for their course. Where approved, these case studies and reports are shared with instructors in similar contexts seeking examples of how the tool might perform with their students.

Implementation Studies

Implementation studies codify use cases and provide additional evidence of the relationship between a use case and the outcomes achieved. Studies are repeated for a variety of use cases and contexts to build evidence of effectiveness that is more relevant to more instructors, and provide wider insights for improvement of the tool and future studies.

An implementation study typically spans a full semester. Quantitative and qualitative data are collected from instructors and students at various time points to accurately codify use of the product in a specific, real-world setting. Descriptive methods are used to document the implementation and instructor and learner perceptions of the product. Correlations are examined between the use of the tool and learner outcomes including engagement, academic performance, retention, and completion.

Findings guide ongoing product optimization and communications with instructors interested in understanding how the tool may impact learning in their setting. Results also support future experimental designs where instructors are selected within their use case to ensure the studies are not biased based on varying implementations.

Rapid-cycle Evaluations

Rapid-cycle evaluations (RCEs) evaluate specific components of a product to understand whether they contribute incrementally to making a product more effective for instructors and learners. In this way, RCEs provide evidence of effectiveness that can help explain results of future impact studies.

The length of a particular RCE varies depending on the research question. For example, analyses using historical data will be more rapid than analyses requiring the collection of new data. A convenience sample (i.e any known user of the product rather than a stratified sample) is used as RCEs are conducted within individual institutions. Like implementation studies, RCEs are replicated in various institutions to increase the representativeness of the population of users.

Use of RCE findings are based on the research questions and the design of the evaluation. For example, incremental effectiveness of a feature or functionality is evaluated when a capability is made available to a randomly selected set of students and similar data are captured on a like peer group. In the absence of a comparison group, effectiveness of a feature or capability

is evaluated by describing perceptions, learner outcomes, and the relationships between use and outcomes. Findings are used to make decisions about refining and optimizing features and to provide context to explain results of future efficacy studies.

Efficacy Studies

Efficacy studies attempt to isolate the impact of a tool on learner outcomes from other contributing factors. By randomly assigning students to a treatment or control group, confounding factors are significantly reduced and any difference in learner outcomes observed between the treatment and control group are assumed to be attributed to the learning tool.

An efficacy study lasts the duration of a course. Only instructors who have engaged in an implementation study are recruited to participate in an efficacy study, so their implementation has already been documented and is understood. All students in the class who agree to participate are considered the population of interest and are randomized into either the treatment or control group within the course. In the case where in-class randomization is not possible, students in a course section using the product are matched to students in a similar section not using the product through a matching algorithm.

> Only instructors and institutions who have engaged in an implementation study are recruited to participate in an efficacy studies…

Data are collected from instructors and all students in both the treatment and control groups at various time points throughout the course. Descriptive statistics and correlations from students who used the product measure its effectiveness. Learner outcomes such as academic performance, retention, and completion are compared between the treatment and control groups to measure the impact of the product.

DATA COLLECTION

Quantitative and qualitative data are collected for each study conducted. Specific data depend on the study being conducted, the research questions being asked, and requirements from the institutional IRB approval. Typically, data collected include:

Student Pre-Survey

An online survey that captures information on self-reported prior academic performance, experience, perception and level of comfort using digital learning tools, perception of a specific discipline course, motivation, and classroom activity challenges.

Instructor Pre-Survey

An online survey that captures information on background and experience teaching higher education and teaching the specific discipline course, time spent preparing for class, experience, perception, and level of comfort using digital learning tools, current challenges with classroom activities, expected implementation of the digital learning tool being studied and other digital learning tools in use, and their expectations of the digital learning tool being studied.

Classroom Observation

Classroom environment and a class are observed and implementation methods are meticulously documented. Student engagement is documented using an observation protocol adapted from Lane & Harris, 2015. For this protocol, a set of students are selected to be observed closely (students are unaware of selection). Classroom activities are

recorded and during each activity the number of students engaged at that time is recorded. Proxies of engagement include active listening, related writing, related reading, appropriate computer use, appropriate student interaction, and appropriate interaction with instructor. The number of students in that set that are actively disengaged are also recorded. Proxies for disengagement include settling in and packing up, being unresponsive, being off-task, inappropriate computer use, inappropriate student interaction, being distracted by another student, and inappropriate interaction with instructor. These records are quantified to establish an engagement metric.

Student Focus Groups

A set of representative students is recruited based on their responses to the pre-survey to take part in a focus group. Focus group protocols are developed based on the product being studied and to probe responses provided on the pre-survey.

Instructor Interview

Interview protocols vary depending on product and implementation, perception of effectiveness, and probes specific to that instructor based on their responses to the pre-survey and the classroom observation.

Student Post-Survey

An online survey that asks students to respond to scales that are parallel to the pre-survey around: perception of and comfort with technology, perception of and comfort with the discipline course being studied, and motivation.

Instructor Post-Survey

An online survey that asks instructors to respond to scales that are parallel to the pre-survey around: perception of, and comfort with digital learning tools, and time taken to prepare for the class being studied. Instructors are also asked to describe any challenges experienced in the course, their implementation of the product, their perception of the product and its impact on their class.

Product Data

Data are extracted from the platform of the product being tested to examine implementation, usage behaviors, user segmentations, and item performance.

Student Records

Instructors provide student records of attendance and academic performance.

Evidence collected throughout the lifecycle of a product creates a portfolio of insights about a product's validity, effectiveness, and impact on student outcomes. The evidence begins with foundational educational research underpinning the product design, explores use cases of the product through implementation studies, and uses rapid-cycle evaluations guide optimization of the product and instructor support. Learning analytics provides a wide range of complementary insights. The portfolio is rounded off with experimental or quasi-experimental impact evaluation studies that use insights and context provided from previous implementation studies and rapid-cycle evaluations. Each of these studies offer practical and actionable insights and a continual feedback loop - to support instructors and students achieving their best outcomes using a product, and ongoing improvement of the product and support.

“Evidence collected throughout the lifecycle of a product creates a portfolio of insights about a product's validity, effectiveness, and impact on student outcomes.”

Limitations

As discussed at the start of this White Paper, measuring the effectiveness and impact of digital learning tools is fundamentally difficult and faces many practical challenges. At Macmillan Learning, we acknowledge that our approach has limitations. We aim to be transparent about these limitations as we continually refine our efforts to design, develop, measure, and optimize products that help students, instructors, and institutions to achieve their best outcomes.

The most critical limitation that any educational researcher faces is the ability to isolate the impact of a product on learner outcomes given it functions as one part of a complex educational ecosystem, and that ecosystem varies by course, instructor, and institution. Isolating the impact of a product in this context using tightly controlled trials may not be practical, timely, or even possible. For these reasons, the framework we propose focuses on examining results in a wide set of local contexts that more closely represent the situations of instructors and other college staff making decisions.

Research is also challenged by digital products that are continually improving and evolving to meet new customer needs, but where instructors need timely insights to support their decisions on adoption and best usage. Measureable learning also takes time and although our framework is structured to produce early insights in parallel with longitudinal studies, those insights are confined to the environments and contexts in which they are gleaned.

Finally, while our framework is ambitious and comprehensive, we acknowledge that the findings of many of our studies, particularly the formative studies, are descriptive and correlational. We are careful that the claims we make about our products based on study results are appropriate for the design, data collection, and analysis and that they are communicated to instructors, learners, and institutions with transparency.

We attempt to mitigate limitations through careful sampling, design, and statistical controls, but there is more work to be done and we look forward to continual improvement. We appreciate that these challenges and the discussions they provoke provide a rich opportunity for us to continue to engage with and contribute to educational research on methods for researching and evaluating digital learning tools.

Ongoing Refinement

There is much to learn about novel, agile approaches to measuring the effectiveness and researching the impact of digital learning tools on learner outcomes, and Macmillan Learning respects unbiased third-party reviews of our approach, studies, and the claims made based on study findings. To provide that, we have formed an **Impact Research Advisory Council** comprising leading experts in designing and measuring the impact of educational technology, measuring effectiveness in practical ways, modeling and evaluating learning performance, standards for measurement in education, data security, and existing and emerging legislation to protect the privacy of human subjects. To ensure that we are continually receiving varied perspectives, Council membership rotates every two years.

The Council has provided guidance and critical feedback as we developed this framework and has reviewed and critiqued our study designs prior to them being implemented. They also continually review our findings and any claims we would like to make for soundness. As we refine our approach, we continue to look to the Council for guidance.

We also recognize that the needs of educators are fluid, so we try to maintain regular communication with instructors and institutions using our products. As part of research, we continually solicit feedback from instructors about the questions they would most like to answer, if the insights we are providing are helping them and their students to be more successful, and how we may support them better.

> …these challenges and the discussions they provoke provide a rich opportunity for us to continue to engage with and contribute to educational research…"

We believe that evaluating the effectiveness and measuring the impact of digital learning tools is essential to improving education and student success. And, that we have a duty to help instructors and institutions to better understand what will work for them and their students, why, and how. We believe in sharing what we know and learning from others and therefore regularly participate in academic conferences, symposiums, and meetings to engage with the broader educational community and transparently share our successes and our challenges.

Conclusion

At Macmillan Learning we embrace the diversity of the students, instructors, and institutions we serve as users of our products and the complexity of the educational ecosystems in which they operate. We work tirelessly to ensure that our products are designed to be highly usable within those ecosystems and that they enhance each instructor's approach and each student's chances of success. However, we are committed to continually evaluating the effectiveness of our products and researching how they influence learner outcomes in order to continually improve and provide insights about effective use.

We believe deeply in the ability of learning to change lives. With such high stakes, we encourage institutions, instructors, and students to demand more transparent, reliable, and relevant evidence so they can make the best informed decisions about what learning products to use, why, and how. To that end, we do not believe that vague or unsubstantiated claims should have a role. Nor do we believe that isolated statements of efficacy are the most meaningful way to help decision makers. Instead, we propose a holistic, evolving, and connected approach that starts by understanding the variety of ways students, instructors, and institutions choose to use products and their local contexts, and progresses through increasingly rigorous studies, repeated across different use cases and environments. We believe this will provide students and educators with more relevant, reliable, and timely insights into if and how a product will be effective and in what circumstances.

We hope that the insights gleaned from our framework contribute to improving learning and are of interest and value to all everyone involved in education.

Works Cited

Lane, E. S., & Harris, S. E. (2015). A new tool for measuring student behavioral engagement in large university classes. Journal of College Science Teaching, 44(6), 83- 91.

U.S. Department of Education, Office of Educational Technology, Expanding Evidence Approaches for Learning in a Digital World, Washington, D.C., 2013.

About the Authors

Dr. Kara McWilliams
Senior Director Impact Research

Kara is passionate about researching the impact of digital technologies in higher education, and how insights can inform teaching and learning. She has ten years of experience conducting qualitative and quantitative investigations of how course and classroom interventions can improve learner outcomes and influence learning gains. She holds a doctorate in Educational Research, Measurement and Evaluation and a master's degree in Curriculum & Instruction from Boston College.

Dr. Adam Black
Chief Learning Officer

Adam is a recognized pioneer in improving learner outcomes. From identifying promising areas of learning science, to directing the development of market-leading digital products (used by more than 26 million learners), and spearheading novel approaches for assessing impact, Adam has 24 years of experience. Adam holds a BSc in Physics from the University of Edinburgh and a PhD in Astrophysics from the University of Cambridge, has two patents pending in analytics innovation, and has won national and global awards for digital product innovation.

Dr. Jeff Bergin
Vice President Learning Research & Design

Jeff has led curriculum, instructional, and learner experience design for various educational technology companies for the past 20 years. Dr. Bergin leads a team of learning researchers and experience designers. In previous roles, he has led the design of personalized and mobile products. Dr. Bergin holds a Ph.D. from Arizona State University and has presented and published on topics including learning design, online learning, and technology-augmented instruction.

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 predicting learners at risk in science using a novel 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.

About Macmillan Learning

Macmillan Learning improves lives through learning. Our legacy of excellence in education continues to inform our approach to developing world-class content with pioneering, interactive tools. Through deep partnership with the world's best researchers, educators, administrators, and developers, we facilitate teaching and learning opportunities that spark student engagement and improve outcomes. 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 **http://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 mining and Impact Research. To learn more about this approach, please visit **http://www.macmillanlearning.com/catalog/page/learningscience**

macmillan learning