MACMILLAN LEARNING RESEARCH METHODS REPORT

MINIMIZING RISK, MAXIMIZING UTILITY

A new method for evaluating educational technology products in alpha

Kara McWilliams PhD, Vice President Impact Research, Macmillan Learning

Billie-Jo Grant PhD, Senior Impact Research Scientist

Ethics and peer review. This research complied with APA ethical standards for research. It was approved by a third-party Institutional Review Board (IRB) prior to recruitment, and then approved by individual institutional IRBs at each participating institution where required. This paper and the results herein, have been peer reviewed, revised, and approved for publication as a Macmillan Technical Report by the Impact Research Advisory Council, a panel of experts in applied research, measurement, educational technology, and the learning sciences.

macmillan learning

Contents

Foreword

At Macmillan Learning we are committed to providing our instructors and students with practical, actionable, and timely insights derived from studies that meet standards for educational and psychological testing. Our goal is to improve teaching and learning by enabling evidence-based decision making and to contribute to the methods and outcome research on digital learning tools. To that end, we take a comprehensive approach to measuring the effectiveness and efficacy of the digital learning tools we produce. Beginning in development, and continuing through use at scale, we partner with instructors and students to conduct studies that are appropriate for the tool's stage in the development lifecycle. Each study contributes unique and increasingly rigorous evidence to the validity and efficacy argument of that tool. Studies also produce insights into usage and engagement patterns among educational contexts that instructors might consider implementing in their own courses. This report represents one study that makes up the larger body of Achieve efficacy research. We are confident in this approach but acknowledge that measuring efficacy is complex, and we are always learning. The authors of this report, and the impact research team as a whole, welcome any comments or feedback on this report or our approach to measuring efficacy.

Kara McWilliams PhD, Vice President Impact Research, Macmillan Learning

Acknowledgements

The work in this report reflects the efforts of many people beyond the authors. We appreciate the commitment that Macmillan Learning has made to ensuring that the tools we are developing are meeting the needs of instructors and students, and improving higher education. We thank our Chief Executive Officer, Ken Michaels, for his continued support on this journey. This study would not have been a success without the partnership of Gareth Hancock and Tim Flem's Product Team, Chelsea Valentine's Technology Team, Alicia Nachman's User Experience team, and Susan Winslow's Editorial group. To each of the leaders at Macmillan and their talented teams, we thank you for your close partnership and willingness to use the insights from research for evidence-based decision making.

We also thank our colleagues on the Learning Research team at Macmillan Learning for their help defining the constructs that we should be measuring in each rapid-cycle evaluation in this study and for their review of the instruments used for data collection. And, our colleagues on the Impact Research team for early reviews of this manuscript. We also thank Pegi Goodman for the design of this report.

In an effort to offer timely, peer reviewed insights to instructors, we are grateful to the Impact Research Advisory Council for their peer review of this report. Their guidance and critique since we began developing our approach to efficacy, ongoing insight throughout each study, and honest reviews of findings have been invaluable. Chris Dede, Michael Feldstien, Sara Finney, Suzanne Lane, Thanos Patelis, and Elana Zieda we are indebted to you.

Most importantly, we want to thank the instructors who partnered with us on this study. This was the first time any of the 38 instructors had participated in a formative evaluation that occurred outside of and independent from their live courses. Without their participation and feedback we would not have been able to gain such deep insight into supporting the optimization of Achieve to enable an effective experience for instructors in various educational contexts. Further, it was their feedback that helped Macmillan Learning make an informed decision about whether Achieve was advanced enough in the development lifecycle that researching its effectiveness in live classrooms with students would facilitate a positive experience.

Abstract

Educational technology has the potential to vastly improve teaching and learning in higher education. Rapid advancements in technology and the pace at which new tools are being developed are leading to alpha products being tested with students in live classrooms, which can have a negative impact on teaching and learning. There are many constraints of educational technology that is early in development, and student success in higher education is critically important. Therefore, it is necessary to develop alternative methods to requiring students to use an alpha version of a learning tool as their primary course material for a full semester. This paper discusses a new method for conducting a formative evaluation of digital learning tools. The method has proven to enable relevant, timely, and actionable insights on product optimization, implementation patterns, and professional development—without requiring the use of an alpha product in a high-stakes environment. A case study of a formative evaluation of new digital platform, Achieve, is used to illustrate the approach. The formative evaluation of the alpha version of Achieve was a longitudinal study conducted with a set of instructors teaching a course in which the solution might be used, but the study was conducted outside of and independent from their live classrooms. The evaluation was comprised of eight unique rapid-cycle evaluations, each lasting one week, that simulated the arc of a traditional semester. Qualitative and quantitative data were collected to develop insights into platform use, implementation patterns, perception, and expectations. Results from the evaluation were used for real-time remediation and optimization of the tool, to understand instructors' chosen implementation patterns, and to inform professional development for future users. Real-time results implemented by the development teams provided confidence among researchers and instructors that a beta version of the platforms used in live classrooms in subsequent studies would not adversely impact the student and instructor experience, but rather contribute positively to important learner outcomes.

Introduction

Early stage testing is a core component of the digital product development lifecycle. However, the designs and methods of early stage testing approaches vary widely based on the development of the product and stakeholder(s) and end-user needs (Ardito et al., 2013; Heiskari & Lehtola, 2009). Generelly, developers work to create a minimum viable product—a solution that is complete enough to introduce to customers, but early enough in development that tester feedback can inform iterative design—and conduct early stage testing on that version of the product (Contigiani & Levinthal, 2018). End-user participation in early stage testing is key to obtaining the appropriate information to ensure the product has the greatest impact when used at scale. These methods are particularly true of developers who embrace a user-centered design[1] approach to product development (Still & Crane, 2017).

Two common early stage testing approaches include alpha and beta testing. The key differences between alpha and beta tests are the availability, robustness, and stability of product features and functionalities, the timeframes of the study, participants in the study, and expected uses of study results. Alpha tests are typically implemented early in development; are generally conducted in-house; and are used to uncover system bugs, identify gaps in the holistic program, and generate feedback for product optimization prior to beta (Awa, 2010).

[1]: User-centered design (UCD) is an iterative approach to software development that is grounded in empathy for the user. It requires that the user interface, and therefore the design of the rest of the system, are the result of an evidence-based synthesis of the end-users' needs and implementation strategies. Throughout a UCS process, a product undergoes ongoing evaluation where users work with designers to ideate solutions, consult on prototypes, and use the product at all stages while researchers and developers observe their behaviors and collect feedback.

Beta tests are implemented once a product has been improved based on alpha test findings and typically shortly before going to market. They are typically conducted with end-users in their contextual environment and generate feedback on implementation patterns, technical challenges when used in context, and optimization insights (Zhu, 2010). In some fields, alpha and beta tests are conducted with end-users because the perspective of the target end-user is specific to the point that it cannot easily be simulated by in-house teams. In these cases the participants are not part of the development process directly, they are providing data that development teams use to make product decisions.

Because of the complex context of higher education, testing with end-users as early in development as possible is critical to creating a tool that will positively impact instructor and learner outcomes (Bhuiyan, 2011; Che Ku Nuraini et al., 2014). Developers and their colleagues sometimes lack the contextual perspective needed to evaluate how an alpha version of a product will perform or whether it will be accepted by students and instructors when used in the classroom. However, testing alpha products with live end-users in higher education is complex. The high stakes nature of higher education introduces methodological and pedagogical risks to testing, which may negatively impact student and instructor success.

Researchers conducting alpha tests in live classrooms face many design and methodology challenges that may bias the validity and reliability of the study results. A reliable study of digital learning tools engages a representative sample of the target population being studied, including end-users of various backgrounds and experience with technology. However, especially because alpha versions of products are more likely to present bugs and create technical difficulties, instructors who are comfortable with technology and who have used it in the past are more likely to agree to participate in the study, skewing the results. Relatedly, because developers will want to minimize disruption to a classroom environment and limit adverse impacts on a student's learning experience with early-stage testing, vendors often provide support to the end-user that is not representative of a live experience. For example, one-to-one onboarding and course set-up support and concierge support may be offered to every instructor throughout the study. "White glove" technical and/or customer support during an alpha test will skew perceptions of ease of use.

Testing products in the alpha stage of development in live classrooms introduces educational challenges as well. Already over-taxed educators have to learn new software, become familiar with the expected limitations of an early iteration of a tool, and accept any workarounds that may be required to achieve a desired outcome. Additionally, although supports are generally provided, instructors and students may experience a disruption in the classroom resulting from a system that is not yet fully built or that presents technical challenges. For example, if a tool presents a bug that limits access, students may miss a homework due date or not be able to review for an upcoming assessment, adversely impacting their learning experience—and perhaps adversely impacting teacher evaluations.

In the absence of even directional evidence that the use of the new tool may lead to positive student outcomes, the risk of in-context testing may not be worth the outcome. Nevertheless, engaging with end-users early and often in the design and development process of digital learning tools is the best way to evaluate and iterate a product. The objective of the study presented in this paper was to design and validate an alternative to alpha testing in live classrooms that, although not tested in-context, maintains the integrity of the evaluation and increases the validity of findings. The methodology was intended to produce results that can be used for immediate product improvement, to begin to understand chosen implementation patterns, and to glean exploratory evidence of learner effectiveness—offering data-driven confidence that the product is ready to be beta-tested with students in a live classroom.

The formative evaluation of the alpha version of the product presented in this study was longitudinal and conducted with a sample of instructors teaching a course that the solution might be used in. However, the study was conducted outside of and independent from their live classroom. The study was conducted over an eight-week period from September to November 2018, it simulated the arc of a semester and allowed instructors to experience the product from both the instructor and student perspectives. Participants engaged in reviews, performance-tasks, and assessments so researchers could gather holistic feedback. Multiple data sources enabled triangulation of results to increase the validity of the inferences made from the findings. And real-time feedback loops were established between instructors, researchers, and developers—allowing for ongoing product optimization and directional evidence of effectiveness.

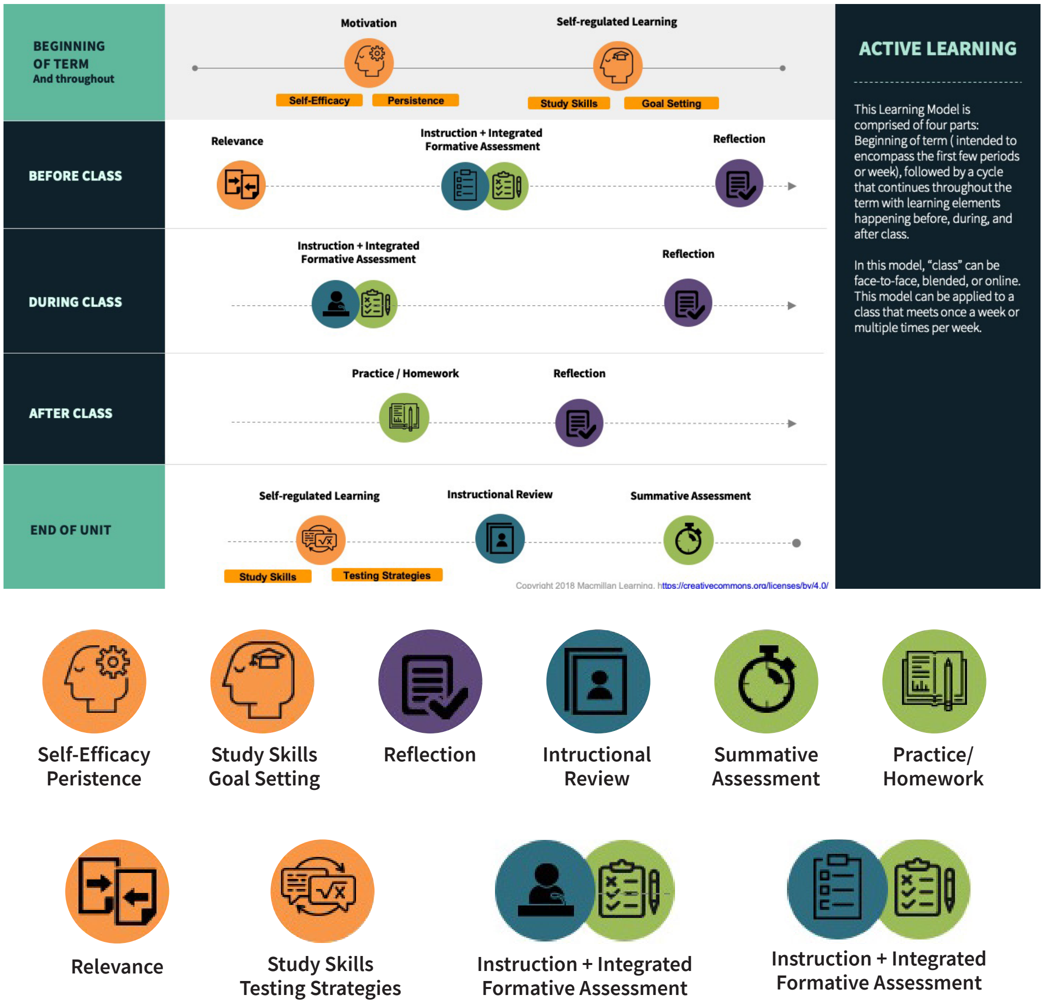
Achieve

The product studied in this formative evaluation was Achieve. Achieve is a digital learning solution developed for higher education courses (at the time of this report publication Achieve was being studied in five disciples: Biology, Calculus, Chemistry, Composition, and Economics). Achieve provides a connected suite of course tools designed to give instructors choice and provide flexible recommendations for optimal pedagogical structures based on the learning sciences. The key principles that Achieve is built on include: everyone has the potential to learn, each learner starts at a different place and learns at their own pace, cognition can be enhanced through technology, an instructor's pedagogy matters, learning is a social activity, and students should be empowered to manage their learning.

Achieve was conceived based on six learning design principles as well as a series of robust learning science foundations that support active learning, objective-driven instruction, formative assessment, and actionable analytics. And, it has been optimized based on the findings of research conducted in close partnership with instructors and students.

A pedagogical structure developed to promote active learning acted as a blueprint for the choreography of Achieve. The model provides an end-to-end structured course that increases instructor efficiencies and supports student success. The active learning model has built in opportunities to support student outcomes beyond course instruction and assessment—like motivation, self-regulated learning, relevance, and study skills. The

Image 1. The learning model underpinning Achieve



ACTIVE LEARNING

This Learning Model is comprised of four parts: Beginning of term (intended to encompass the first few periods or week), followed by a cycle that continues throughout the term with learning elements happening before, during, and after class.

In this model, “class” can be face-to-face, blended, or online. This model can be applied to a class that meets once a week or multiple times per week.

Copyright 2018 Macmillan Learning. https://creativecommons.org/licenses/by/4.0/

active learning model also enables metacognition by providing preflection and reflection activities that prompt evaluation of developing knowledge. And, a host of proven instructional content—such as publisher-provided materials, lecture slides, and instructional reviews—offer opportunities to review or provide new learning-objective aligned instructional information. Integrated formative assessment, practice activities and homework, and end-of-unit or term summative assessments provide an ongoing assessment of learning and feedback for increased learning

Procedures

This research complied with American Psychological Association ethical standards for research. It was approved by a third-party Institutional Review Board (IRB) prior to participant recruitment.

PARTICIPANTS. In total, 38 instructors teaching at a total of 36 institutions from various cities across the United States were recruited to participate in the study.

Information collected on the baseline survey enabled an examination of instructor characteristics. Most instructors in the sample teach Economics (63%), and the largest proportion (45%) have been teaching in this discipline for more than 15 years. The majority of instructors (84%) have used a publisher-provided digital learning tool in the past, and a substantial proportion (42%) reported being extremely comfortable using digital technology in the classroom and strongly agreed (55%) that publisher-provided digital learning tools can enhance classroom pedagogy. Because the recruited instructors tended to have a positive perception of digital learning tools used in the classroom prior to the study, researchers disaggregated the data and compared findings between the set of instructors with a positive perception and those with a negative perception of the effectiveness of digital learning tools.

RECRUITMENT PROCEDURES. Instructors were recruited by researchers. During recruitment instructors were given a high-level introduction to Achieve and an outline of the specific arc of activities they would be completing and evaluating each week, but by design formal training was not provided. When instructors agreed to participate, they signed consent forms and memoranda of understanding of their responsibilities as participants.

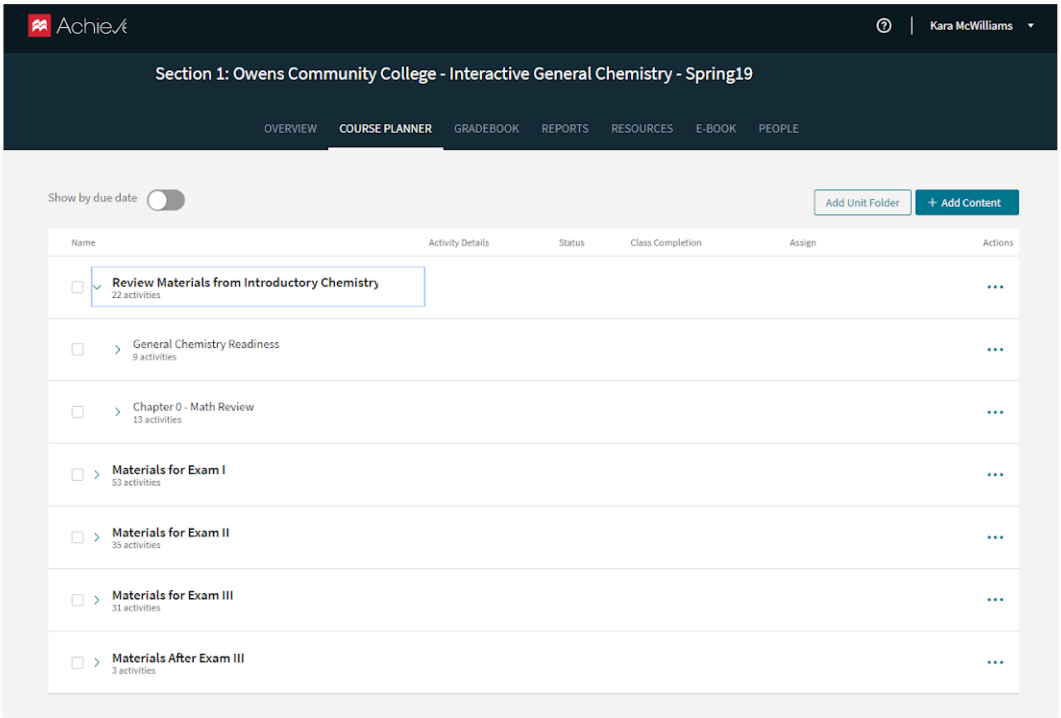
Methods and analyses

The formative evaluation of Achieve was conducted with instructors who teach relevant courses but conducted outside of and independent from an instructor's live course. Each Monday morning, instructors were emailed a description of a unique activity for evaluating one of the core components of Achieve, instructions on how to complete the activity, and instruments to evaluate the experience. Activities were due back to researchers by the following Sunday to allow for immediate analysis of results and feedback to developers, editors, and other stakeholder groups. The eight core features of Achieve to be evaluated were agreed upon by internal stakeholders, and instructions and instruments were developed by researchers. An overview of the eight-week arc of rapid-cycle evaluations is presented in Table 2 and detailed implementation, description, and analytical procedures follow the table.

Table 1: Arc of eight week rapid-cycle evaluations

Week	Feature evaluated	Method	Use of results
1	Course set-up	Instructors use the course planner to independently set up their courses like they would during a live semester. Decisions made while setting up their course are journaled and instructors complete a survey about perception	• Evaluation of which activities are being assigned to inform content development • Evaluation of which activities are being assigned to inform content development • Usability and user experience improvements
2	Diagnostic assessment	Instructors complete a diagnostic assessment from the perspective of a student. They work through the pre-test, study plan, and post-test. They journal their expected use, complete a perception survey and platform data are analyzed to understand usage.	• Evaluation of instructor expected use to inform content development • Evaluation of perception of assessments to inform improvement planning • Navigation, usability and user experience improvements
3	Diagnostic analytics	Instructors view the instructor-facing diagnostic dashboard report, they complete a performance task and a perception survey	• Performance task provides valid measure of understanding to inform clarification of metrics or descriptions • Perception data to validate performance task results and inform improvement planning
4	Student experience	Instructors complete a set of assignments (varying activity types) from the perspective of a student, they complete a perception survey and platform data are pulled to investigate usage patterns and performance.	• Evaluation of instructor expected use of activity types they might not have been familiar with and to inform content development • Evaluation of perception of different activity types to inform improvement planning • Navigation, usability and user experience improvements
5	Dashboard reports	Instructors view the complete instructor-facing dashboard reports, they complete a performance task and a perception survey.	*Performance task provides valid measure of understanding to inform clarification of metrics or descriptions *Perception data to validate performance task results and inform improvement planning
6	Active learning	Instructors review the pre-populated questions that are designed to be used with iClicker, a student response system, they provide feedback on each item, journal their expected use, and complete a perception survey.	• Evaluation of whether instructors would use a student response system if they had pre-developed items; inform professional development • Content expertise to support optimization of items
7	Gradebook	Instructors evaluate a gradebook that has been pre-populated with synthetic data, they complete a performance task and a perception survey.	• Performance task provides valid measure of understanding to inform clarification of metrics or descriptions • Perception data to validate performance task results and inform improvement planning
8	Course set-up	Instructors are given a fresh course template. Instructors use the course planner to independently set up their courses like they would during a live semester. They journal their decisions in setting up their course and complete a survey about perception	• Alignment study between initial course set up and final week course set up, did changes results from becoming more familiar with activity types - to inform training plans, professional development, and in-platform support planning • Change in perception to inform whether ease of use improves over time

Week One: Course set-up



Section 1: Owens Community College - Interactive General Chemistry - Spring19

	Name	Activity Details	Status	Class Completion	Assign	Actions
□	Review Materials from Introductory Chemistry 22 activities					...
□	General Chemistry Readiness 9 activities					...
□	Chapter 0 - Math Review 13 activities					...
□	Materials for Exam I 53 activities					...
□	Materials for Exam II 35 activities					...
□	Materials for Exam III 31 activities					...
□	Materials After Exam III 3 activities					...

Image 2: Early design of Achieve course planner

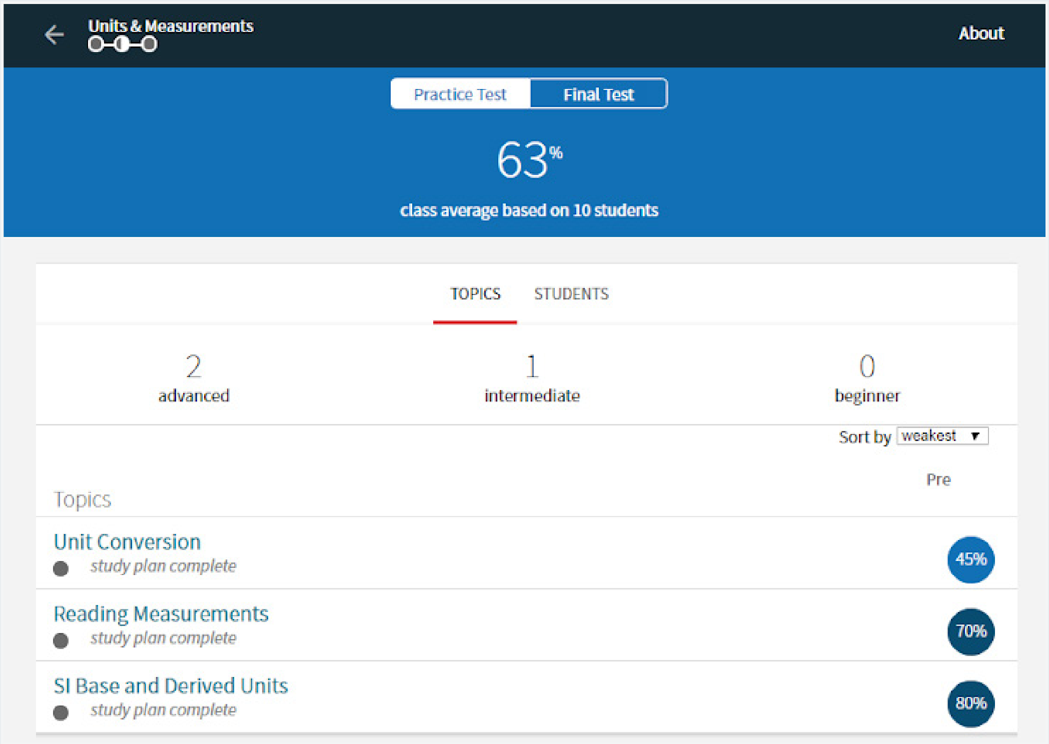
WEEK 1: *Course set-up.* The course planner within Achieve was developed to support a course choreography based on learning science. Each searchable resource was tagged by learning objective and tagged with a recommendation of whether it should be assigned pre-class, during-class, or post-class. Course set-up, however, is flexible; so instructors can search, assign, and tag as they choose. The rapid-cycle evaluation was designed to measure: the extent to which instructors utilized the tools within the course planner tool, the usability of the system, their perception of the course planner tool, and the extent to which instructors set up their courses as Achieve stakeholders expected.

Instructors were sent an email on Monday that read, in part, “This week we are interested to see the choices you would make when setting up your course. We are asking you to review the resources available for two chapters, set the chapters up like you would in a live course, and comment on the process. Instructions for completing the activity are included in this email. Please complete the activity by Sunday evening. This week you should have to commit no more than two hours to this evaluation.”

Following the course set-up, instructors were asked to complete a survey that measured their perception of the ease of use, usability, efficiency, and quality of the course set-up process.

Following the submission of their activities, three analyses were conducted. First, researchers compared the course each instructor had set up with a template course that had been built by stakeholders and aligned the similarities and differences. Then, surveys were individually analyzed in an effort to qualify why the course had been set up in that way. Any divergence from expected course set-up that could not be explained by survey data was noted on an individual instructor interview protocol form to probe during the final interview. Finally, survey data were analyzed to understand the ease of use, usability, and general perception to inform improvement efforts.

Week Two: Diagnostic assessment



63%

class average based on 10 students

2 advanced | 1 intermediate | 0 beginner

Topics

- Unit Conversion ● study plan complete 45%
- Reading Measurements ● study plan complete 70%
- SI Base and Derived Units ● study plan complete 80%

Image 3: Early design of student diagnostic assessment

WEEK 2: *Diagnostic assessment.* Achieve includes a diagnostic tool that is made up of a pre-assessment, an individualized study plan, a post-assessment, a student-facing analytics dashboard, and an instructor-facing analytics dashboard. The intent of the diagnostic tool is to support the identification and remediation of skills gaps early in the semester. The diagnostic tool also provides instructors with early insights into skills gaps their students may have and the extent to which those gaps were closed through the use of the study plan. The second rapid-cycle evaluation was conducted by instructors from a student perspective and was designed to measure: instructor's perception of the diagnostic tool, the extent to which they perceive that the tool would be used by their students, the extent to which the content aligned with their curriculum, and their perception of ease of use and usability.

Instructors were sent an email on Monday morning that read, in part, "*This week you will be completing one of the diagnostic portions of General Chemistry Readiness (for the Expressions section of the chapter). This is a "pre-test" to help students and their instructors understand the mastery of some core competencies students need for this section as well as a study plan to help them remediate any gaps that were identified in the pre-test, and concludes with a "final test" to help students see their growth. Please complete the diagnostic assessment by Sunday evening. This week you should have to commit no more than one to one and one half hours to this evaluation. You can break up the time you spend on the activity but please complete the activity logs (15-20 minute survey) in one sitting.*" Once instructors had finished completing the diagnostic assessment from the perspective of a student, they were asked to complete two activity logs.

Activity log one was a survey that measured the intuitiveness of the student-facing analytics dashboard. The items included in this survey had a right or wrong answer. For example, instructors were asked, "How many resources were included in your personalized study plan?" In response, instructors could either write in a number or choose "I do not know". The results of these items helped validly measure intuitiveness rather than instructors' perceptions of intuitiveness. Activity log two was a survey that measured perception, asking instructors questions like "Rate the extent to which you agree that your students would find value in the diagnostic tool". Activity logs one and two were analyzed to understand accuracy of responses and perception. Again, any anomalies observed were tracked on the individual instructor interview protocol for probing at the end of the semester.

WEEK THREE: Instructor diagnostic analytics dashboards

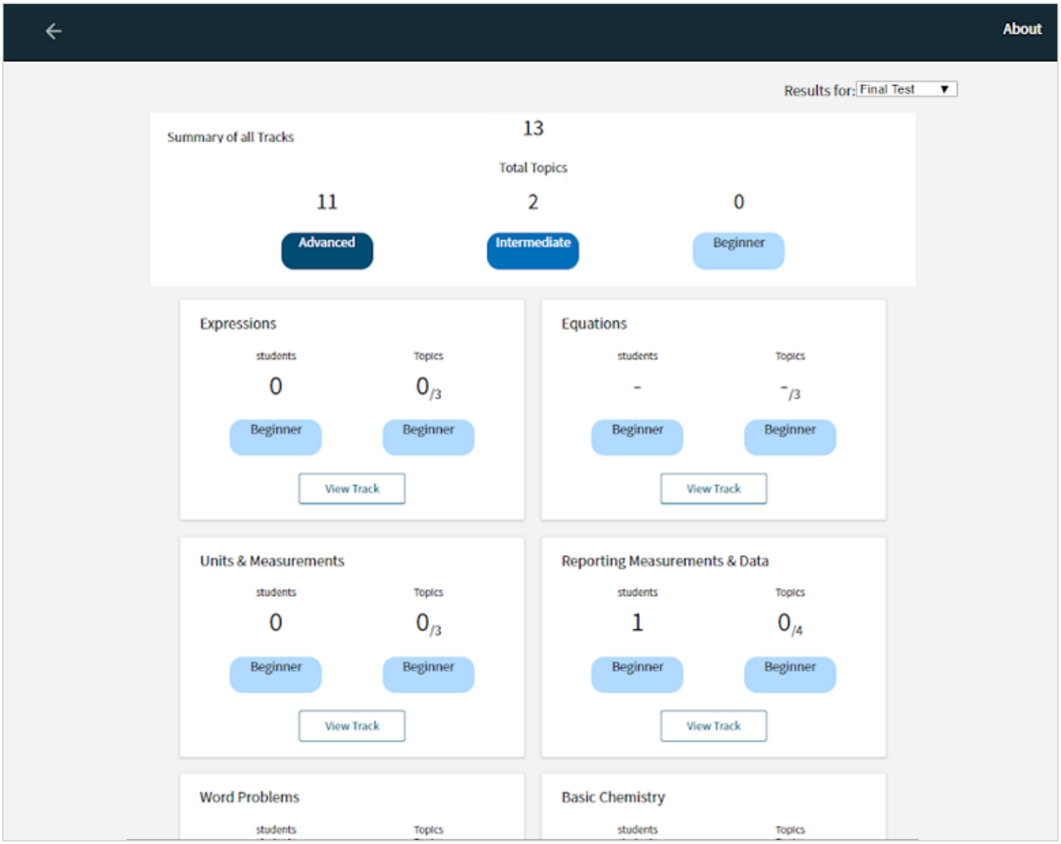


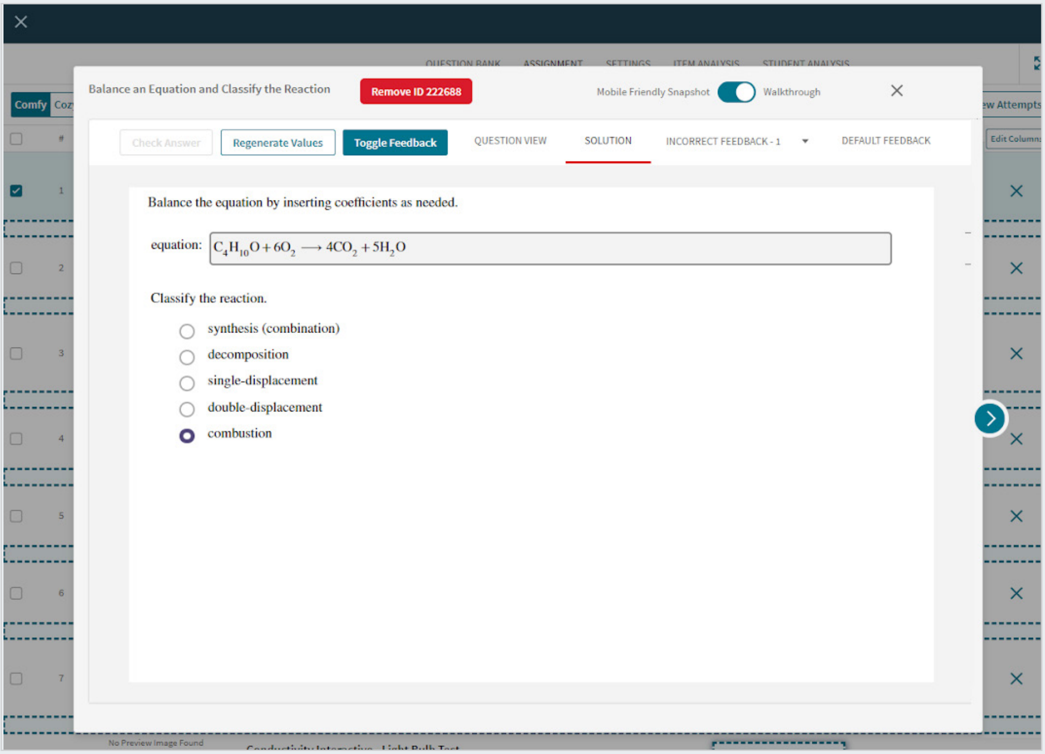
Image 4: Early design of instructor diagnostic dashboard

WEEK 3: *Instructor diagnostic analytics dashboards.* As noted, the diagnostic assessment includes an instructor-facing dashboard intended to offer instructors insights into how academically prepared their students are and whether there are skills gaps to close early in the semester. The third rapid-cycle evaluation was conducted by instructors from an instructor's perspective and was designed to measure: instructor's perception of the diagnostic analytics dashboard, the extent to which instructors could identify components of the dashboard and interpret them, how instructors expected that they would use the information from the diagnostic dashboards, and the ease of use and usability of the analytics dashboards.

Instructors were sent an email Monday morning that read, in part, "*Last week you completed the student experience in the General Chemistry Readiness diagnostic pre-test. That was a view from the student's perspective. This week you will be reviewing the dashboard analytics that instructors see once students complete the diagnostics. We are interested in your feedback on whether the information provided in the dashboards is intuitive as well as useful, therefore we are asking that you complete an activity as well as an activity log (two surveys). Please complete the activity by Sunday evening. This week you should have to commit no more than one and one half hours to this evaluation. Please try to complete this activity in one sitting.*"

Instructors were sent a pre-populated course with simulated student data included to represent an active course. Activity log one was a survey that measured the intuitiveness of the instructor-facing analytics dashboard. The items included in this survey had a right or wrong answer. For example, instructors were asked, "List the students in your course who have gaps in biochemistry skills." Instructors were asked to write the individual student names or select "I do not know". The results of these items helped validly measure intuitiveness rather than instructors' perceptions of intuitiveness. Activity log two was a survey that measured perception, asking instructors questions like "Rate the extent to which you would use the analytics presented here to intervene with students. If yes, please describe how you would intervene based on these analytics". Activity logs one and two were analyzed to understand accuracy of responses and perception. Again, any anomalies observed were tracked on the individual instructor interview protocol for probing at the end of the semester.

WEEK FOUR: The general student experience.



Balance the equation by inserting coefficients as needed.

equation: $C_4H_{10}O + 6O_2 \longrightarrow 4CO_2 + 5H_2O$

Classify the reaction.

○ synthesis (combination)
○ decomposition
○ single-displacement
○ double-displacement
● combustion

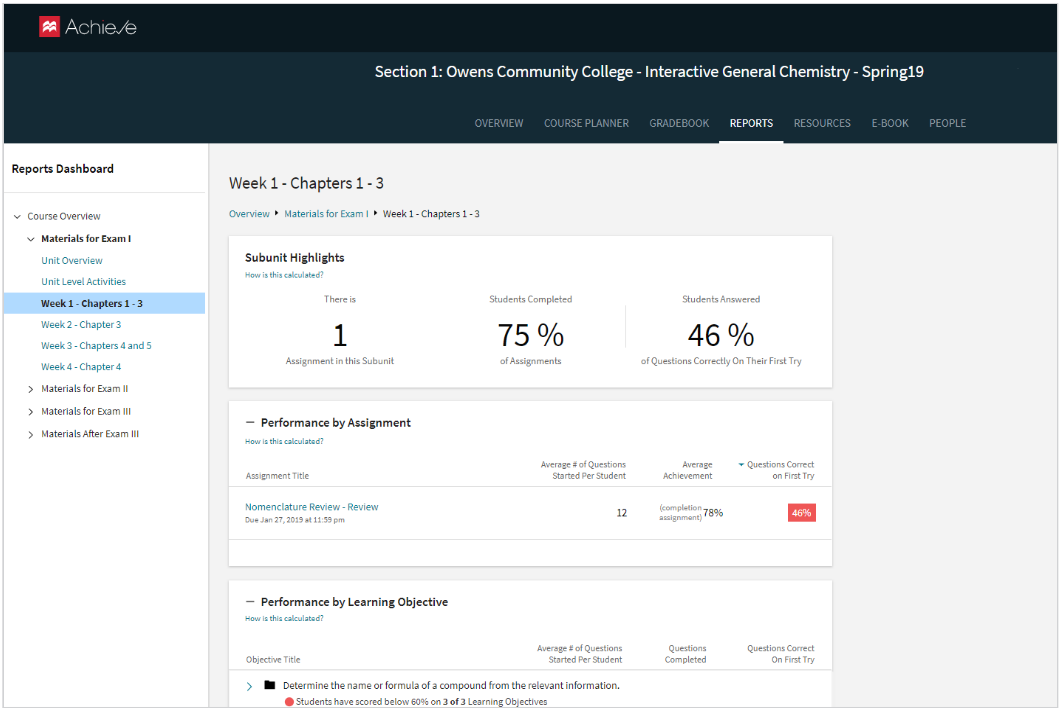
Image 5: Example student assessment

WEEK 4: *The general student experience.* In the fourth rapid-cycle evaluation, instructors were provided with a curated course that a student might see. The course provided an example of each of the resources available to instructors to assign. Instructors were asked to work through the chapter as if they were a student. They were asked to complete all activities and to simulate the expected motivation level of a student completing the activities for credit. Taken together, the activities assigned each semester are expected to motivate students to come to class prepared to participate, actively engage in class discussion, review effectively, and achieve learning gains. The evaluation in week four was intended to measure: instructor's perception of the content covered in student activities and assessments, the quality of the activities and assessments, perceived student acceptance and level of engagement in the activities and assessments, and perception of the ease of use, usability, and overall quality.

Instructors were sent an email Monday morning that read, in part, "*This week you will be reviewing Chapter 7 of General Chemistry from the perspective of a student. We are providing you with student credentials to a curated course that probably looks different from the course you set up earlier in this evaluation, we are only asking you to work through a subset of the chapter. Instructions for completing the activity are included in this email. Please complete the activity by Sunday evening. This week you should have to commit no more than two and a half hours to this evaluation. You can break up the time you spend on the activity but please complete the activity log (15-20 minute survey) in one sitting.*"

Data were analyzed in three ways in week four. First, data were extracted from Achieve's platform to understand how instructors engaged in Achieve that week and to provide context for the perception of responses in the survey. Then, survey data were analyzed to understand instructor perceptions of the quality of the content they reviewed, how engaging the activities were, how challenging the assessments were, and general usability and user experience. Finally, performance on the activities and assessments were examined to investigate the level of challenge in the assessments. Anomalies or questions were recorded in the individual instructor interview protocols.

WEEK FIVE: General dashboards and just-in-time teaching



Section 1: Owens Community College - Interactive General Chemistry - Spring19

Week 1 - Chapters 1 - 3

Subunit Highlights

There is	Students Completed	Students Answered
1 Assignment in this Subunit	75 % of Assignments	46 % of Questions Correctly On Their First Try

Performance by Assignment

Assignment Title	Average # of Questions Started Per Student	Average Achievement	Questions Correct on First Try
Nomenclature Review - Review Due Jan 27, 2019 at 11:59 pm	12	(completion assignment) 78%	46%

Performance by Learning Objective

Objective Title	Average # of Questions Started Per Student	Questions Completed	Questions Correct On First Try
Determine the name or formula of a compound from the relevant information. Students have scored below 60% on **3 of 3** Learning Objectives			

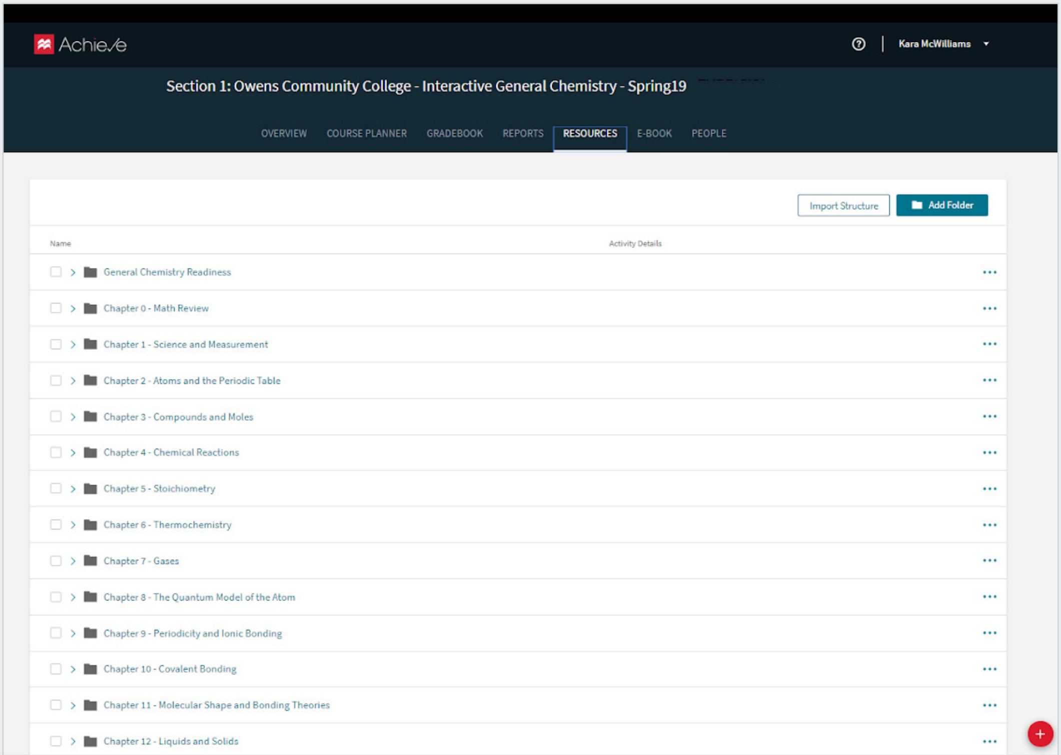
Image 6: Early design of Achieve instructor-facing dashboard

WEEK FIVE: *General dashboards and just-in-time teaching.* In the fifth rapid-cycle evaluation, instructors were asked to evaluate the general analytics dashboards. The dashboards provide insights into course-level and student-level performance. Insights enable just-in-time teaching so instructors can modify their lecture or in-class activities based on student performance on prior activities or assessments. Dashboards also offer the ability to implement targeted interventions for individual students.

A course was populated with simulated student data resembling a true course (i.e., students of varying levels of academic preparedness and levels of motivation and persistence). A dashboard representative of a live course emerged based on the simulated course data. Instructors were sent an email on Monday morning that read, in part, *"Last week you reviewed Chapter 7 of Achieve for IGC and GCR from a student perspective. This week you are going to be reviewing the dashboard of Chapter 7 that instructors see. The assignments and activities follow the same organization that you worked through last week but have been populated with synthetic data for the purposes of this activity. Please complete the activity by Sunday evening. This week you should have to commit no more than one and one half hours to this evaluation. Please try to complete this activity in one sitting."*

Similar to the rapid-cycle evaluation of the diagnostics dashboard, there were multiple activities associated with the evaluation of the general dashboards. Because there are multiple "layers" of the dashboard, participants were asked to complete a usability and findability activity where items were similar to, "Locate the section of the dashboard that indicates the proportion of students who have completed homework activity 3. Were you able to locate this section? What proportion turned it in on time? How difficult was it to locate this section?" The second activity measured intuitiveness and asked questions such as, "Which learning objectives are students struggling to comprehend"? The third activity measured perception and expectations and asked questions such as "Please rate the extent to which the insights provided in this dashboard would enable you to modify your in-class time to better meet the needs of your students"? and "How often do you expect that you would review your course dashboards"? Notes that were listed in the individual instructor interview form included things like areas to which an instructor could not successfully navigate, inaccurate definitions, and misinterpretations. We further probed those issues during the end-of-course survey.

WEEK SIX: Active learning in the classroom.



Section 1: Owens Community College - Interactive General Chemistry - Spring19

Name	Activity Details
□ General Chemistry Readiness	•••
□ Chapter 0 - Math Review	•••
□ Chapter 1 - Science and Measurement	•••
□ Chapter 2 - Atoms and the Periodic Table	•••
□ Chapter 3 - Compounds and Moles	•••
□ Chapter 4 - Chemical Reactions	•••
□ Chapter 5 - Stoichiometry	•••
□ Chapter 6 - Thermochemistry	•••
□ Chapter 7 - Gases	•••
□ Chapter 8 - The Quantum Model of the Atom	•••
□ Chapter 9 - Periodicity and Ionic Bonding	•••
□ Chapter 10 - Covalent Bonding	•••
□ Chapter 11 - Molecular Shape and Bonding Theories	•••
□ Chapter 12 - Liquids and Solids	•••

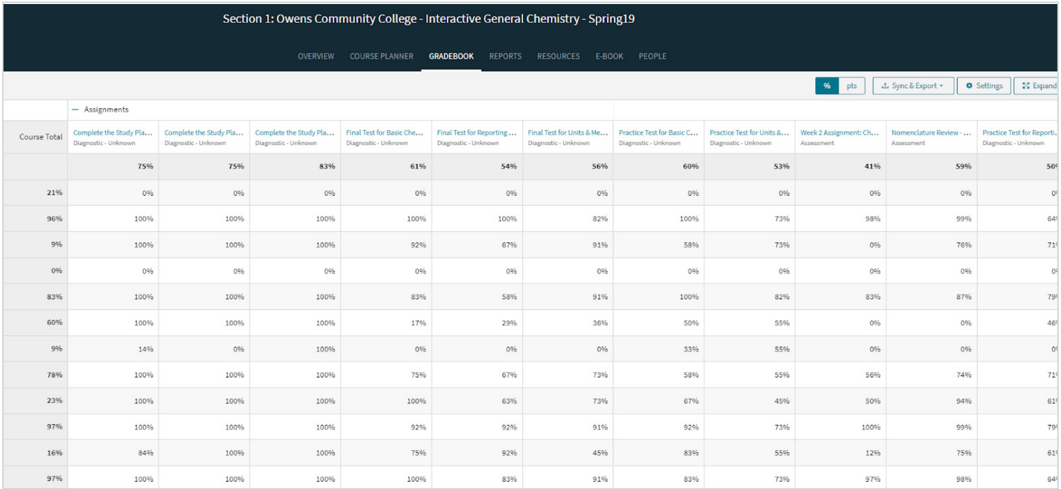
Image 7: Early design of Achieve resource list

WEEK SIX: *Active learning in the classroom.* In the sixth rapid-cycle evaluation, instructors were asked to evaluate the components of Achieve that were developed to promote active learning in the classroom—including content developed to be used with student response systems, static worksheets expected to enable peer-to-peer learning, and case studies expected to promote group-work during in-class time.

An email went out to instructors on Monday morning that read, in part, *"One of our goals with the Achieve full course solution in Chemistry is to offer a structure that supports active learning in the classroom. This week, we would like to get your thoughts on active learning in the classroom in general, and more specifically, review some materials that we suggest will support active learning and give us your thoughts (an in-class worksheet and provided iClicker questions). Please complete the activity by Sunday evening. This week you should have to commit no more than one and one half hours to this evaluation. Please try to complete this activity in one sitting."*

The evaluation was a general review of the content contained within the resources. The weekly log was a survey that captured information on the participant's perception of the quality of the content, their expected use of the product, and their perception of whether use of the resources would support active learning in the classroom. Any areas that required further clarification were noted on the individual instructor interview protocol and probed during interviews.

WEEK SEVEN: Gradebook



Section 1: Owens Community College - Interactive General Chemistry - Spring19

Course Total	Complete the Study Pla… Diagnostic - Unknown	Complete the Study Pla… Diagnostic - Unknown	Complete the Study Pla… Diagnostic - Unknown	Final Test for Basic Che… Diagnostic - Unknown	Final Test for Reporting … Diagnostic - Unknown	Final Test for Units & Me… Diagnostic - Unknown	Practice Test for Basic C… Diagnostic - Unknown	Practice Test for Units &… Diagnostic - Unknown	Week 2 Assignment: Ch… Assessment	Nomenclature Review - … Assessment	Practice Test for Reporti… Diagnostic - Unknown
	75%	75%	83%	61%	54%	56%	60%	53%	41%	59%	50%
21%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
96%	100%	100%	100%	100%	100%	82%	100%	73%	98%	99%	64%
9%	100%	100%	100%	92%	67%	91%	58%	73%	0%	76%	71%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
83%	100%	100%	100%	83%	58%	91%	100%	82%	83%	87%	79%
60%	100%	100%	100%	17%	29%	36%	50%	55%	0%	0%	46%
9%	14%	0%	100%	0%	0%	0%	33%	55%	0%	0%	0%
78%	100%	100%	100%	75%	67%	73%	58%	55%	56%	74%	71%
23%	100%	100%	100%	100%	63%	73%	67%	45%	50%	94%	61%
97%	100%	100%	100%	92%	92%	91%	92%	73%	100%	99%	79%
16%	84%	100%	100%	75%	92%	45%	83%	55%	12%	75%	61%
97%	100%	100%	100%	100%	83%	91%	83%	73%	97%	98%	64%

Image 8: Early design of Achieve gradebook

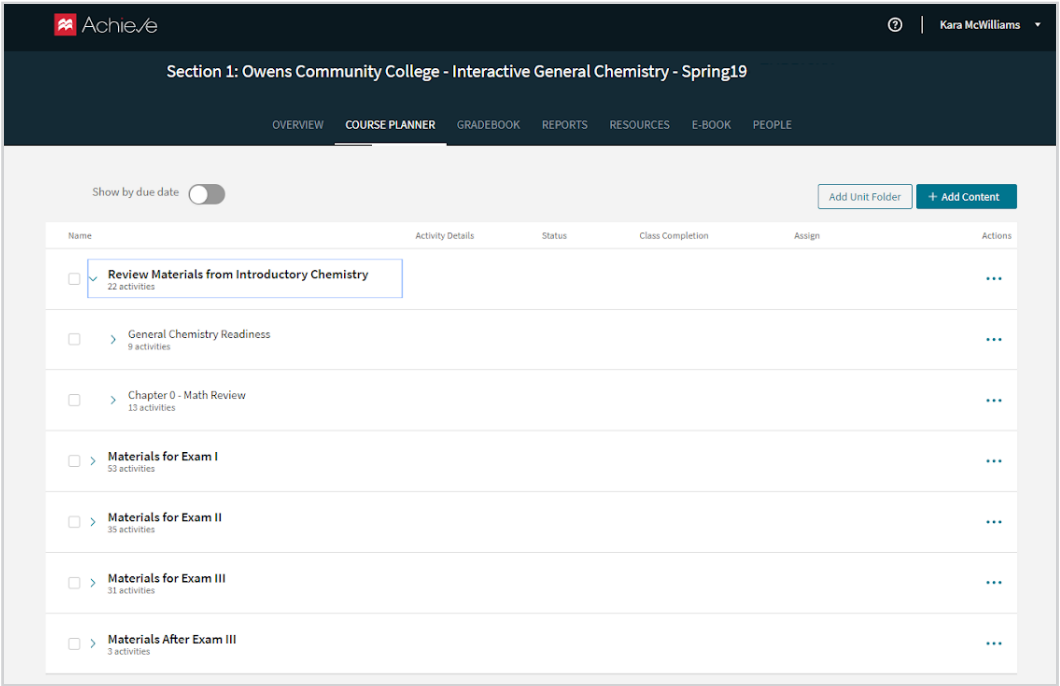
WEEK SEVEN: *The gradebook*. In the seventh rapid-cycle evaluation, the quality, usability, user experience, and intuitiveness of the gradebook in Achieve were measured. Instructors once again accessed a course that had been populated with simulated data so that they viewed a full gradebook.

An email went out to instructors on Monday morning that read, in part, *"This week in the evaluation we would like you to review the current Achieve gradebook. It is populated with synthetic data from the activities you completed in Chapters 7 and 8 (you'll see your own data there too!), please complete the gradebook activity, and fill out the weekly activity log (two survey gizmos). Please complete the activity by Sunday evening. This week you should have to commit no more than two hours to this evaluation. Please try to complete this activity in one sitting."*

The evaluation consisted of a performance assessment and a perception survey. The performance assessment was unmoderated and provided instructors with instructions for navigating around the gradebook. Instructors were asked to complete tasks such as "It is the end of the semester and you want to export your grades out of Achieve, please complete that task". Perception questions asked things like "Please rate the extent to which you agree that the Achieve gradebook is easy to use".

Two analyses were conducted. First, platform data were extracted to observe the paths that instructors took to complete the tasks they were assigned and to understand navigation choices made by instructors. Then, survey data were analyzed to receive feedback on perception and expected use. As always, anomalies were noted in individual instructor interview protocols.

WEEK EIGHT: Reset your course



Section 1: Owens Community College - Interactive General Chemistry - Spring19

Name	Activity Details	Status	Class Completion	Assign	Actions
□ ⌄ Review Materials from Introductory Chemistry 22 activities					···
□ › General Chemistry Readiness 9 activities					···
□ › Chapter 0 - Math Review 13 activities					···
□ › Materials for Exam I 53 activities					···
□ › Materials for Exam II 35 activities					···
□ › Materials for Exam III 31 activities					···
□ › Materials After Exam III 3 activities					···

Image 9: Early design of the Achieve course planner

WEEK EIGHT: *Reset your course.* In the final rapid-cycle evaluation, instructors were given another "empty" course and asked to replicate the activity of setting up their course in the first week of the evaluation. The intention of the rapid-cycle evaluation was to observe whether instructors' course set-up process, or the resulting course, was dissimilar from the first rapid-cycle evaluation. We hypothesized that after using the platform for some time, instructors would assign different activities and/or their navigation between activity types would become more efficient. These insights could inform initial training or ongoing professional development efforts.

An email went out to instructors on Monday morning that read, in part, *"Now that you have experienced many of the different components of Achieve for General Chemistry, we'd like to see if you would set your course up any differently than you did in week 1. Please visit your course page and set up your course for Chapters 3 and 4 again. Please complete the activity by Sunday evening. This week you should have to commit no more than one and one half hours to this evaluation."* Instructors completed a perception survey after the activity.

Three analyses were conducted. First, the rebuilt course was compared with the course that instructors built in the beginning of the semester; and an alignment study was conducted. Then, the rebuilt course was compared to the template course created by stakeholders to measure alignment to the expected course setup. Finally, survey data were analyzed to understand whether their perception of the course set-up process changed over time since the first week of the evaluation.

Final interview. Following the final rapid-cycle evaluation, each instructor was contacted to set up a one-hour virtual interview. Each interview followed the same protocol of five questions, and then the individual instructor interview protocols were delivered. Probes were posed in real time, and the final question asked instructors to share any other information about their experience in Achieve that would help us with development and optimization.

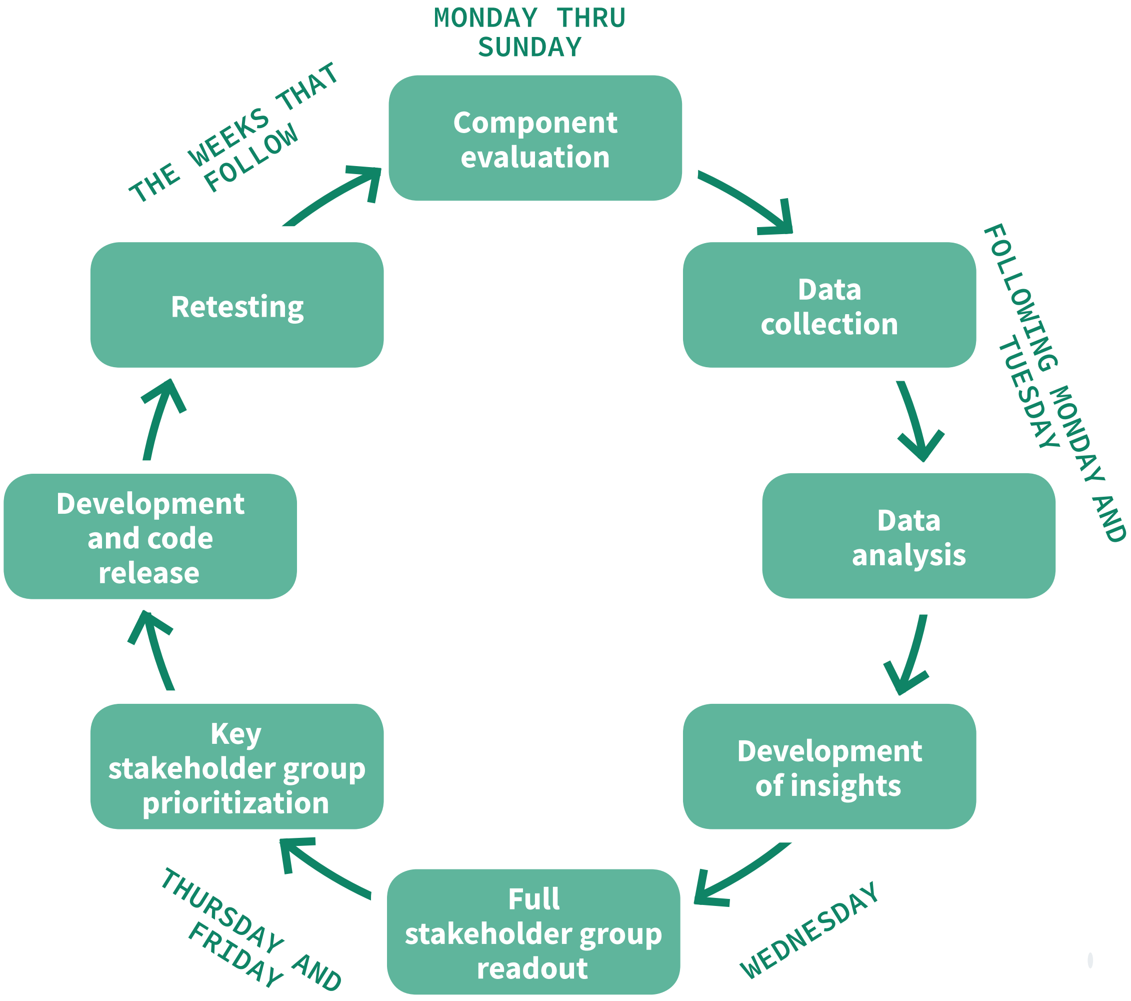
Insight loops

The majority of the key findings in this evaluation are specific to product features and beyond the scope of this methods paper. Nevertheless, some key findings are presented in this section so the reader can see how the data were actioned.

Ability to evaluate a tool out of context. On the post-survey, instructors were asked to rate their level of agreement (scale 1 = "strongly disagree" through 4 = "strongly agree") with a set of statements meant to investigate whether they perceived that they could evaluate the alpha version of Achieve effectively using the method of the formative evaluation. We hypothesized that even though this evaluation was not being conducted outside of their live classroom, the methodology would enable valid evaluation.

Instructors tended to agree that they were able to validly evaluate Achieve through this methodology. Average responses from the individual questions are presented in Figure 5.

Figure 1. Weekly insight loop



1. *Component evaluation.* A feature was released to instructors on a Monday morning and they were asked to complete one or more tasks to evaluate that feature by the following Sunday night.

2. *Data collection.* Instructor data are accessed Monday morning through one or more surveys and aggregated. Usage data were extracted to evaluate access, progress, completion, and level of engagement.

3. *Data analysis.* Perception data were analyzed overall and by institutional and instructor subgroups to measure whether there was variability among educational contexts. Performance data were analyzed for accuracy. Data were also disaggregated by discipline to understand whether there were different needs by discipline.

4. *Development of insights.* A brief slide deck was created highlighting the top insights derived from the rapid-cycle evaluation conducted that week. The deck contained specific data points, illustrations to highlight the insights derived from the data, instructor quotes from open response items to contextualize insights, and specific recommendations for optimization.

5. *Full stakeholder-group readout.* Each week, a live readout was delivered to any interested stakeholders participating in the design and development of Achieve. During the readout key findings were presented and stakeholders had the opportunity to ask clarifying questions and discuss interesting insights that emerged among a cross-functional group. During the readout stakeholders were invited to suggest any modifications to upcoming weeks evaluations based on design or development changes that had occurred in the previous week. Full stakeholder group readouts included an average of 57 live participants (range among weeks 33 to 81).

6. *Key stakeholder group prioritization.* Following the full stakeholder group readout a smaller group convened to action the insights presented in the readout. This smaller, more focused meeting included one key members from Learning Science, Product, Engineering, Editorial, Customer Experience, and Marketing. Here, the insights were prioritized for immediate work or added to a backlog.

7. *Development and code release.* Insights that were prioritized for immediate work were developed and releases deployed to production (the digital environment in which instructors were testing).

8. *Re-evaluation*. If a release was deployed within the formative evaluation window, instructors were asked to comment on the optimizations made and whether they now met expectations and needs.

Key findings

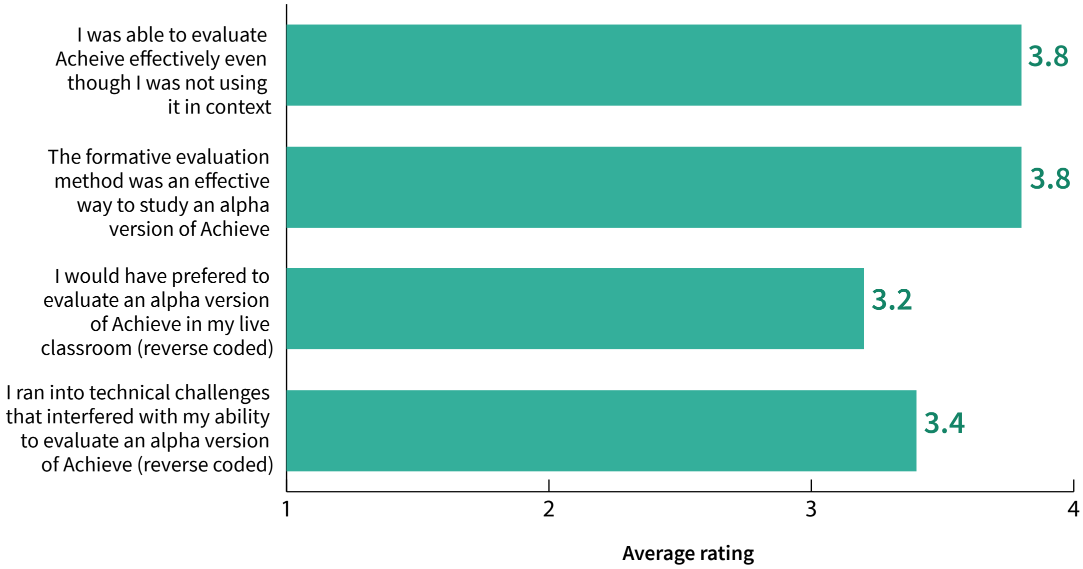
The majority of the key findings in this evaluation are specific to product features and beyond the scope of this methods paper. Nevertheless, some key findings are presented in this section so the reader can see how the data were actioned.

Ability to evaluate a tool out of context. On the post-survey, instructors were asked to rate their level of agreement (scale 1 = "strongly disagree" through 4 = "strongly agree") with a set of statements meant to investigate whether they perceived that they could evaluate the alpha version of Achieve effectively using the method of the formative evaluation. We hypothesized that even though this evaluation was not being conducted outside of their live classroom, the methodology would enable valid evaluation.

Instructors tended to agree that they were able to validly evaluate Achieve through this methodology. Average responses from the individual questions are presented in Figure 5.

> "I would spend my class time differently, spend more time on what they didn't understand. I would assign something else. If they got it, I would not cover it. I wouldn't waste my time if they already got it."

Figure 2. Average instructor responses to a set of items measuring perception of validity of evaluation

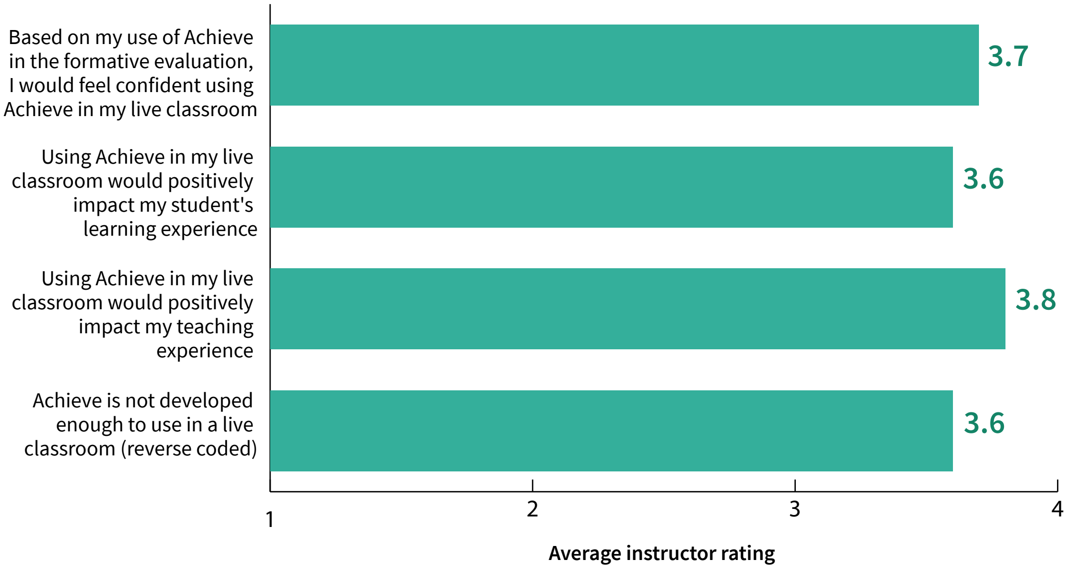


Usability of the alpha version of Achieve. Instructors were asked to respond to the items that make up the System Usability Scale[2] score, and the overall SUS score was 79 (an average score of a fully released digital product is 69). The instructor SUS was measured based on instructors' holistic review of the system over eight weeks. Note that in many cases the instructors were provided guidance on how to navigate the system to get to the section where the activity was located, confounding any measure of "findability" that was measured in the scale. So, the results are suggestive of early usability. Instructors were also asked to respond to one question that asked them to rate the extent to which they agreed that Achieve was easy to use (scale 1 = "strongly disagree" through 4 = "strongly agree") and the average rating was 3.1.

Perception of the alpha version of Achieve. The majority of instructors in the sample (72%) reported that adopting ACHIEVE would increase their teaching efficiencies; a qualitative analysis of responses indicated that this would primarily be a result of all resources being in one system, tags assigned by Macmillan Learning making identification and assignment of resource more efficient (e.g. learning objectives, learning path designations), and information in the dashboard analytics enabling just-in-time teaching and more efficient use of in-class time.

[2] Brooke, J. (1986). System Usability Scale. Digital Equipment Corporation.

Figure 3. Measure of instructor confidence using Achieve in their live classroom



> “I would use the information from the pre-class assignments to figure out where students are struggling. It is really helpful to have system identify the challenges for you.”

Almost all instructors reported that adopting ACHIEVE would increase their students’ efficiencies; a qualitative analysis of responses indicated that this would primarily be a result of all resources being in one system, the ability to identify concepts that students had already mastered and areas of gaps to efficiently target their lectures and activities, and an instructor’s ability to personalize feedback based on individualized student reporting.

Confidence in using a beta version in a live classroom. Instructors were asked to respond to a set of items evaluating their confidence using Achieve in a live classroom with their students. The majority of instructors (91%) agreed that they would feel confident using Achieve in their classroom. Results from all items can be found in Figure 6.

Key optimizations

Each feature evaluated as part of the formative evaluation has been optimized—and in many cases re-evaluated—based on instructor feedback. Instructors in the study provided critical feedback about what features and functionality were table stakes and must be further developed before they could effectively use it with their students. For example, the feedback received on the gradebook was that it was not far enough along in development and its current form risked inefficiencies for instructors and possible misrepresentations of grade data. As a result, gradebook optimizations were prioritized and improvements made early enough for the formative evaluation participants to re-evaluate and validate that the improved gradebook met their needs.

Given capacity and bandwidth, not all recommendations for optimization could be developed before the following semester where Achieve was expected to be studied in live classrooms. To support prioritization we asked instructors to identify the table stakes improvements that should be focused on and which could be included in the back-log for release in subsequent semesters. Instructor prioritization was weighted heavily in product development roadmapping decisions.

Key optimizations of four of the most critical pieces instructors identified: course planner, the student experience, dashboards, and gradebook are presented below.

COURSE PLANNER OPTIMIZATIONS BASED ON INSTRUCTOR FEEDBACK:

Addition of left-side navigation bar

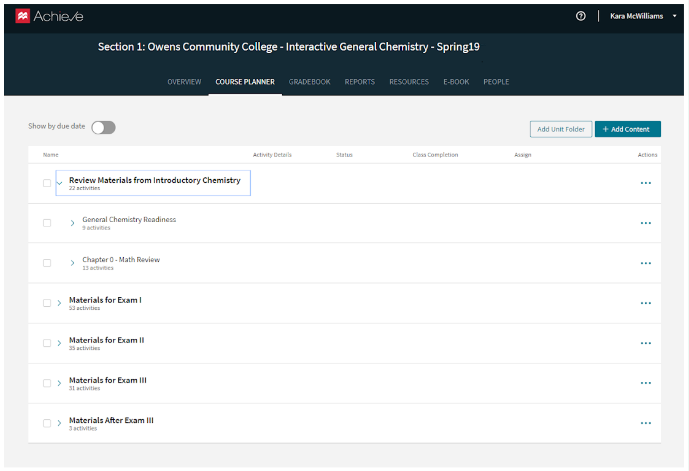
Options to filter resources based on key tags like content type and learning path

Inclusion of icons to more easily recognize tabs

Functionality improvements to assignment management

Added more interactive item types to assessments

Course planner evaluated in formative evaluation

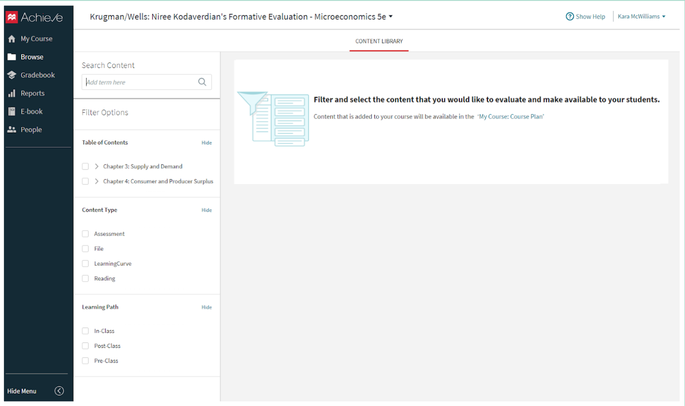


Section 1: Owens Community College - Interactive General Chemistry - Spring19

Name	Activity Details	Status	Class Completion	Assign	Actions
□ Review Materials from Introductory Chemistry 22 activities					...
□ General Chemistry Readiness 9 activities					...
□ Chapter 0 - Math Review 13 activities					...
□ Materials for Exam I 53 activities					...
□ Materials for Exam II 35 activities					...
□ Materials for Exam III 31 activities					...
□ Materials After Exam III 3 activities					...

Image 10: Early design of Achieve course planner

Course planner redesign based on instructor feedback



Krugman/Wells: Niree Kodaverdian's Formative Evaluation - Microeconomics 5e

Table of Contents

- □ Chapter 3: Supply and Demand
- □ Chapter 4: Consumer and Producer Surplus

Content Type

- □ Assessment
- □ File
- □ LearningCurve
- □ Reading

Learning Path

- □ In-Class
- □ Post-Class
- □ Pre-Class

Filter and select the content that you would like to evaluate and make available to your students.

Content that is added to your course will be available in the 'My Course: Course Plan'

Image 11: Redesigned course planner

STUDENT EXPERIENCE OPTIMIZATIONS BASED ON INSTRUCTOR FEEDBACK

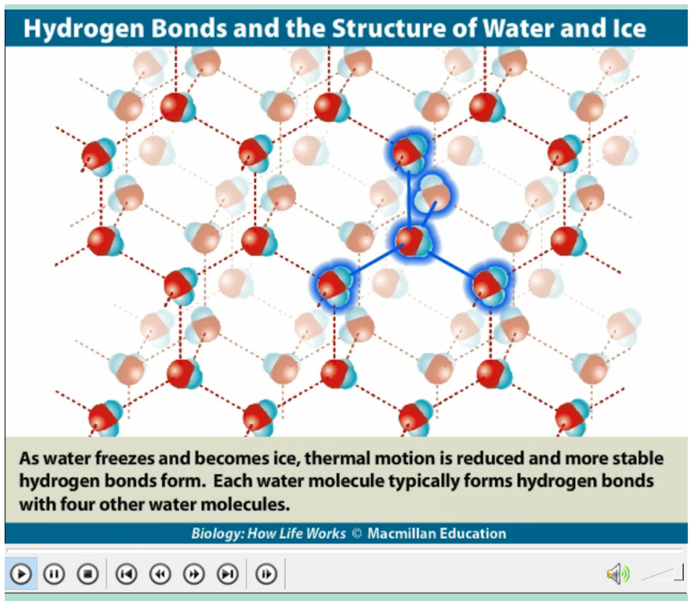
Included additional pre-lecture video tutorials

Included additional video lectures on specific topics instructors report students often find challenging

Added more interactive item types to assessments

Made a number of usability and user experience improvements like navigation, labeling, and functionality of interactive item types

Optimizations to the student experience



Hydrogen Bonds and the Structure of Water and Ice

As water freezes and becomes ice, thermal motion is reduced and more stable hydrogen bonds form. Each water molecule typically forms hydrogen bonds with four other water molecules.

Biology: How Life Works © Macmillan Education

Image 12: Pre-lecture video

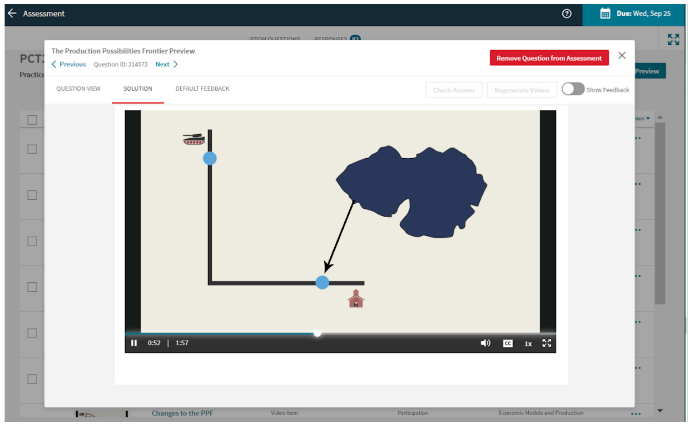
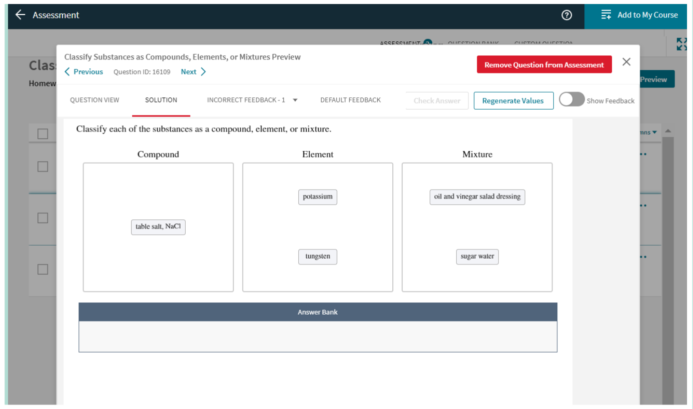


Image 13:Video with assessment questions



Classify each of the substances as a compound, element, or mixture.

Compound	Element	Mixture
table salt, NaCl	potassium	oil and vinegar salad dressing
	tungsten	sugar water

Image 14: interactive assessment questions

DASHBOARD REPORT OPTIMIZATION BASED ON INSTRUCTOR FEEDBACK

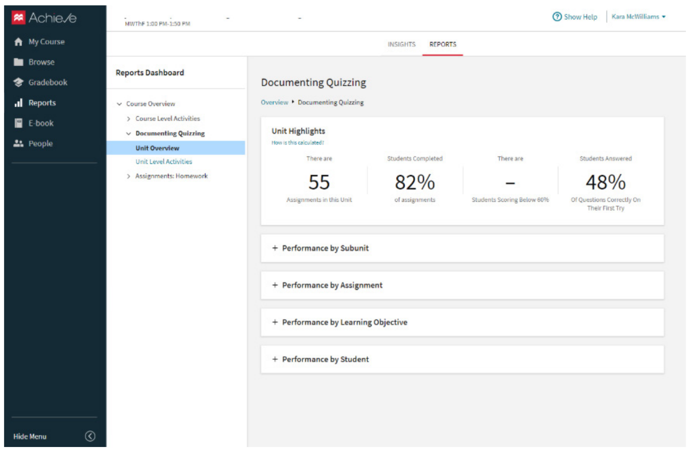
Added left side navigation so instructors can efficiently select the level of the report to review

Created a “top insight” initial screen so instructors can efficiently see which learning objectives from the current unit they might want to review with their students

Included new metrics to offer insights into student engagement and performance

Made a number of user experience design optimizations (example: including color blocking to quickly identify students who are engaging and performing

Dashboard report optimization based on instructor feedback



Documenting Quizzing

Overview ‣ Documenting Quizzing

Unit Highlights

How is this calculated?

There are	Students Completed	There are	Students Answered
55	82%	–	48%
Assignments in this Unit	of assignments	Students Scoring Below 60%	Of Questions Correctly On Their First Try

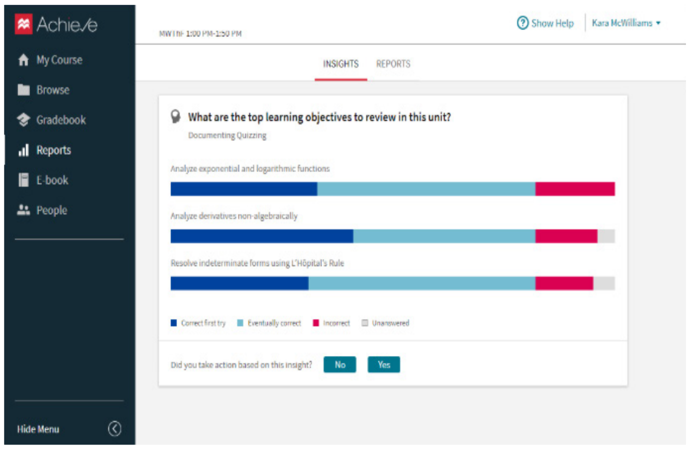
+ Performance by Subunit

+ Performance by Assignment

+ Performance by Learning Objective

+ Performance by Student

Image 15: Redesigned dashboard reports

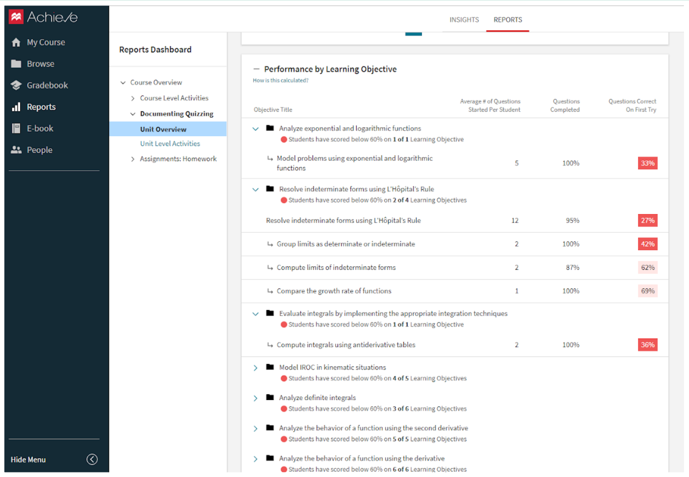


What are the top learning objectives to review in this unit?

Documenting Quizzing

Did you take action based on this insight? No Yes

Image 16: Redesigned dashboard reports



– Performance by Learning Objective

How is this calculated?

Objective Title	Average # of Questions Started Per Student	Questions Completed	Questions Correct On First Try
Analyze exponential and logarithmic functions ● Students have scored below 60% on **1 of 1** Learning Objective			
↳ Model problems using exponential and logarithmic functions	5	100%	33%
Resolve indeterminate forms using L’Hôpital’s Rule ● Students have scored below 60% on **2 of 4** Learning Objectives			
Resolve indeterminate forms using L’Hôpital’s Rule	12	95%	27%
↳ Group limits as determinate or indeterminate	2	100%	42%
↳ Compute limits of indeterminate forms	2	87%	62%
↳ Compare the growth rate of functions	1	100%	69%
Evaluate integrals by implementing the appropriate integration techniques ● Students have scored below 60% on **1 of 1** Learning Objective			
↳ Compute integrals using antiderivative tables	2	100%	36%
Model IROC in kinematic situations ● Students have scored below 60% on **4 of 5** Learning Objectives			
Analyze definite integrals ● Students have scored below 60% on **3 of 6** Learning Objectives			
Analyze the behavior of a function using the second derivative ● Students have scored below 60% on **5 of 5** Learning Objectives			
Analyze the behavior of a function using the derivative ● Students have scored below 60% on **6 of 6** Learning Objectives			

Image 17: Redesigned dashboard reports

GRADEBOOK OPTIMIZATIONS BASED ON INSTRUCTOR FEEDBACK

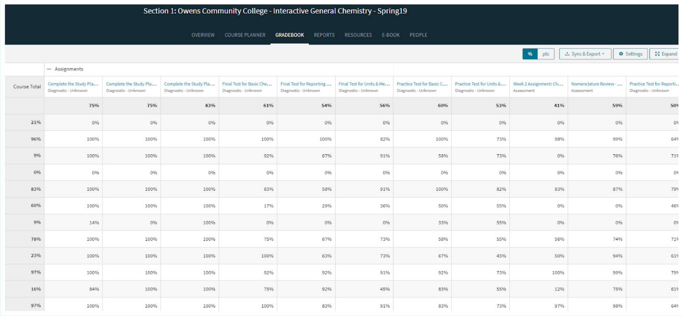
Added horizontal and vertical scroll bars with frozen columns for ease of navigation

Included a highlight function to efficiently link student, activity, and score

Improved LMS integration and export capabilities

Updated calculations to more closely reflect instructor pedagogy

Gradebook evaluated in formative evaluation

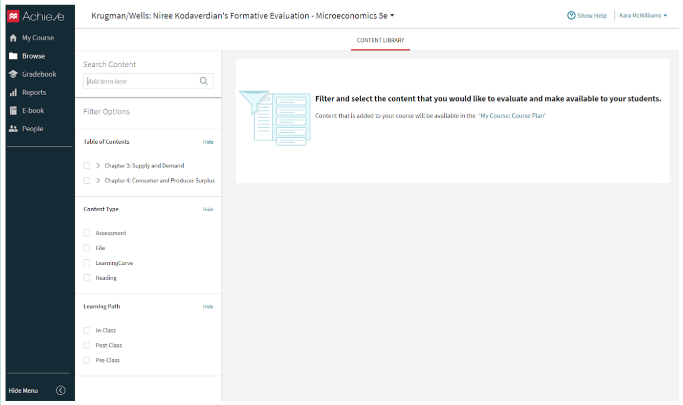


Section 1: Owens Community College - Interactive General Chemistry - Spring19

Course Total	Complete the Study Pla… Diagnostic - Unknown	Complete the Study Pla… Diagnostic - Unknown	Complete the Study Pla… Diagnostic - Unknown	Final Test for Basic Che… Diagnostic - Unknown	Final Test for Reporting … Diagnostic - Unknown	Final Test for Units & Me… Diagnostic - Unknown	Practice Test for Basic C… Diagnostic - Unknown	Practice Test for Units &… Diagnostic - Unknown	Week 2 Assignment: Ch… Assessment	Nomenclature Review - … Assessment	Practice Test for Reporti… Diagnostic - Unknown
	75%	**75%**	**83%**	**61%**	**54%**	**56%**	**60%**	**53%**	**41%**	**59%**	**50%**
21%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
96%	100%	100%	100%	100%	100%	82%	100%	73%	98%	99%	64%
9%	100%	100%	100%	92%	67%	91%	58%	73%	0%	76%	71%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
83%	100%	100%	100%	83%	58%	91%	100%	82%	83%	87%	79%
60%	100%	100%	100%	17%	29%	36%	50%	55%	0%	0%	46%
9%	14%	0%	100%	0%	0%	0%	33%	55%	0%	0%	0%
78%	100%	100%	100%	75%	67%	73%	58%	55%	56%	74%	71%
23%	100%	100%	100%	100%	63%	73%	67%	45%	50%	94%	61%
97%	100%	100%	100%	92%	92%	91%	92%	73%	100%	99%	79%
16%	84%	100%	100%	75%	92%	45%	83%	55%	12%	75%	61%
97%	100%	100%	100%	100%	83%	91%	83%	73%	97%	98%	64%

Image 18: Early design of gradebook

Gradebook optimizations based on instructor feedback



Krugman/Wells: Niree Kodaverdian's Formative Evaluation - Microeconomics 5e

Search Content

Add term here

Filter Options

Table of Contents — Hide

□ Chapter 3: Supply and Demand
□ Chapter 4: Consumer and Producer Surplus

Content Type — Hide

□ Assessment
□ File
□ LearningCurve
□ Reading

Learning Path — Hide

□ In-Class
□ Post-Class
□ Pre-Class

Filter and select the content that you would like to evaluate and make available to your students.

Content that is added to your course will be available in the 'My Course: Course Plan'

Image 19: Redesigned gradebook

Best practices

The formative evaluation of a tool in an alpha form presented in this paper is new to the educational technology industry. Because this was a novel approach, and in an effort to extend the methodology literature on formative evaluation and support researchers who might want to implement the design, a set of best practices that emerged from the study are presented in this section.

1. Practice discipline when identifying the components of the user journey that are key to include in the evaluation. As digital learning tools become more complex, there are many components that stakeholders will suggest would benefit from inclusion in the formative evaluation. If the weekly feature testing becomes too complex, though, the lightweight nature of a rapid-cycle evaluation is lost. Key to a successful formative evaluation is close coordination with key stakeholders to establish the core components of the tool to be evaluated and an understanding of what stakeholder groups want to learn about each component. Then, confirm whether any of the components can be validated through other user testing, like moderated usability studies. Our recommendation is to narrow the components to no more than eight, or overcomplication becomes a risk.

2. Recruit a representative sample of participants. Work with key stakeholders to have a clear understanding of the universe of users of the digital tool being developed. Then stratification should happen at three levels:institutional segment (institution size, type, selectivity, etc.), discipline (Biology, Calculus, Chemistry, etc.), and user segment (years teaching, comfort with technology, etc.). A stratified sample will enable early feedback from the universe of users while keeping the sample size manageable. Participants can be placed in institutional segments based on publicly available information. A useful method for accurately stratifying participants into user segments is constructing a very brief interest survey that is sent out with an initial recruitment message. The survey should be no longer than four to six questions, allowing researchers to easily codify participants into identified user segments.

3. Maintain organization and leverage automation. There are substantial logistical considerations when planning a formative evaluation: activities that range from developing multiple individual accounts for instructors; population of simulated data; weekly communication with participants; collection, aggregation, and management of data; report writing; and feedback. Developing a clear and detailed project plan prior to beginning the evaluation will take substantial time, but it will pay dividends. For example, in the formative evaluation described here, all of the weekly emails describing locations, log-in information, and instructions (both initial and reminder) were developed well in advance of the study. After a full quality control investigation of all emails, messages were scheduled to be sent on the appropriate days. Methods like these cut down on the day-to-day manual logistics that are required in formative evaluations.

4. Capture data beyond perception. In early stage testing, particularly during tool discovery, researchers rely on perception data when making decisions about product optimization and tool effectiveness. During a formative evaluation, move beyond perception data and gather performance and cognitive data. To the extent possible, each activity should include a performance task (e.g. Build a course as you would during a live semester), a set of cognitive questions (e.g. How many students in this course failed to demonstrate mastery of chemical equations?) and perception data (e.g. Rate the extent to which you believe the dashboard analytics provide actionable insights).

> "I've never been part of a review before where I saw my feedback actioned within weeks, it's great"!

5. Establish tight internal feedback loops. The formative evaluation described in this paper was effective because of the tight feedback loops that were established with internal stakeholders. The primary stakeholders were the Product, Technology, Editorial, and User Experience teams. Each Sunday when data were submitted from participants, they were pulled and analyzed; and key insights were fed back to the stakeholders within one week. When attempting to turn valid and reliable data around quickly, it is key to be selective of what needs to be communicated in real time to inform decision making and what could wait for summative reporting. Insights that are key for weekly reporting should be established with the stakeholder group prior to conducting the evaluation, as should the form-factor of how they will receive these data.

6. Provide feedback to study participants. The researchers who conducted this evaluation found that acting on weekly feedback from the study participants was related to more robust insights from them. For example, substantial feedback was provided to add additional metrics to the dashboard analytics. Within four weeks of receiving this feedback, the analytics were added. When the updates were communicated to participants, many commented on how impressed they were that their feedback was being implemented and started providing even more robust comments. Communicating back to participants that their insights are being actioned helps them feel more engaged, which will help make the evaluation more successful.

7. Conduct a structured end-of-study interview. Upon completion of the rapid-cycle evaluations, conducting a structured interview will help provide color to the overall evaluation. A strategy that was employed in this formative evaluation was to maintain an individual instructor interview protocol throughout the evaluations, noting any anomalies or findings that required additional information. For example, an instructor who assigned only homework in both iterations of the course setup evaluation was probed in the interview about why that was. Researchers learned that they took too long to evaluate what each of the other resources were; and if there were clearer descriptions, they probably would have assigned other resources. This discovery led to a workaround making the resource descriptions more obvious to instructors unfamiliar with Macmillan's tools. The individual protocols are valuable because it keeps the interview efficient and the feedback valuable to provide context to specific weekly evaluations.

Conclusion

A great benefit of digital learning tools is the agile development that enables the early release of alpha versions of products and continual iteration and optimization. In education though, testing an alpha version of a product in a live classroom risks negatively impacting a student's learning experience. The formative evaluation presented in this paper is a novel approach to maximizing insights about product validation, optimization, and perceived effectiveness while minimizing risk. Possibly the most beneficial outcome is the ability to evaluate whether the learning tool is at a point in development where it can be used in live courses without risking a negative learning experience for students.

In the formative evaluation of Achieve presented here, we learned that the decision to implement this methodology before beta testing in a live classroom was the right one. There were too many table stake features identified as in need of further development by instructors to have made the instructor and student experience in the live classroom a positive one.

One of the most significant outcomes of this evaluation was the way that cross-functional teams at Macmillan huddled around the study and worked together to incorporate the results into the development of Achieve. The process acts as an important case study for incorporating instructor partnership and evidence-based decisions into the agile process.

Limitations and mitigations

Like with all novel research and evaluation methods, there were limitations to the formative evaluation shared in this methods paper. The main limitation to the formative evaluation is that students are not represented in the evaluation. As noted, this is by design. Developers made the decision that the product was not far enough in maturity to be used in a live classroom without risking a negative learning experience. The limitation was mitigated by conducting separate evaluations with student groups outside of their live coursework. Students were sent components of Achieve to complete an evaluation, so early feedback could be gleaned from that important group of users. Other testing was conducted with students as well, including usability and user experience testing.

Another limitation included the study being conducted out of context; so in many cases, participants were asked to comment on their expected use of features rather than observing their actual use. This limitation could not be mitigated in this study. In a subsequent study, however, some participants used Achieve in their live courses; and researchers were able to validate the extent to which participants' expected behavior aligned to their actual behavior.

Note on data privacy

Prior to data collection, this study and the associated consent forms and instruments were reviewed and approved (found exempt) by the Human Resources Research Organization (HumRRO). HumRRO is a third-party Institutional Review Board organization with no affiliation with Macmillan Learning (federal wide assurance number 00009492 and IRB number 00000257). Macmillan Learning seeks independent and unfunded third-party review to eliminate any bias in decision of exemption. The data collected in this study, which are provided by consenting instructors, are initially identifiable. However, once a random identifier is generated identifiable data are destroyed. Data are provided in secure storage locations, and access is permitted only to the primary investigator in the study. For full details of our data handling and storage privacy procedures, contact Kara McWilliams, Vice President Impact Research at Macmillan Learning at kara.mcwilliams@macmillan.com.

REFERENCES

Ardito, C. et al. (2013) Enabling end users to create, annotate and share personal information spaces. In: Dittrich Y., Burnett M., Mørch A., Redmiles D. (eds) End-user development. IS-EUD 2013. Lecture Notes in Computer Science, 7897. Springer, Berlin, Heidelberg. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-38706-7_5

Awa, H. O. (2010). Democratizing the new product development process: A new dimension of value creation and marketing concept. International Business Research, 3(2), 49. Retrieved from https://www.ccsenet.org/journal/index.php/ibr/article/view/5624

Barab, S. (2014). Design-based research. In R. Sawyer (Ed.), The Cambridge Handbook of the Learning Sciences (Cambridge Handbooks in Psychology, pp. 151-170). Cambridge: Cambridge University Press. Retrieved from https://doi.org/10.1017/CBO9781139519526.011

Bhuiyan, N. (2011). A framework for successful new product development. Journal of Industrial Engineering and Management, 4(4), p. 746-770. Retrieved from http://dx.doi.org/10.3926/jiem.334

Che Ku Nuraini Che Ku Mohd, Faaizah Shahbodin, & Naim Che Pee (2014). Exploring the potential technology in personalized learning environment (PLE). J. Appl. Sci. & Agric., 9(18): 61-65. Retrieved from https://www.semanticscholar.org/paper/Exploring-the-Potential-Technology-in-Personalized-Mohd-Shahbodin/9f0f7b72e21029b28cc9cbe8fdb1fcdd5096e6c1

Contigiani, A., & Levinthal, D. (2018). Situating the construct of lean startup: adjacent ‘conversations’ and possible future directions. SSRN Electronic Journal. Retrieved from http://dx.doi.org/10.2139/ssrn.3174799

Dolan, R., & Matthews, J. (1993). Maximizing the utility of customer product testing: beta test design and management. Journal of Product Innovation Management, 1993; 10(4):318-330. Retrieved from https://doi.org/10.1016/0737-6782(93)90074-Z

Heiskari, J., & Lehtola, L. (2009). Investigating the state of user involvement in practice. Proceedings - Asia-Pacific Software Engineering Conference, APSEC. 433-440. Retrieved from https://ieeexplore.ieee.org/document/5358805

McKenney, S. E., & Reeves, T. C. (2012). Conducting educational design research. Milton Park, Abingdon, Oxon: Routledge.

Still, B., & Crane, K. (2017). Fundamentals of user-centered design: A practical approach. CRC Press: Boca Raton, Florida.

Tullis, T., & Albert, B. (2008). Measuring the user experience: Collecting, analyzing, and presenting usability metrics. Amsterdam: Elsevier/Morgan Kaufmann.

Zhu, Z. (2010). Study on beta testing of web application. The 2nd International Conference on Computer and Automation Engineering (ICCAE), Singapore, 2010, pp. 423-426. Retrieved from https://ieeexplore.ieee.org/document/5451922

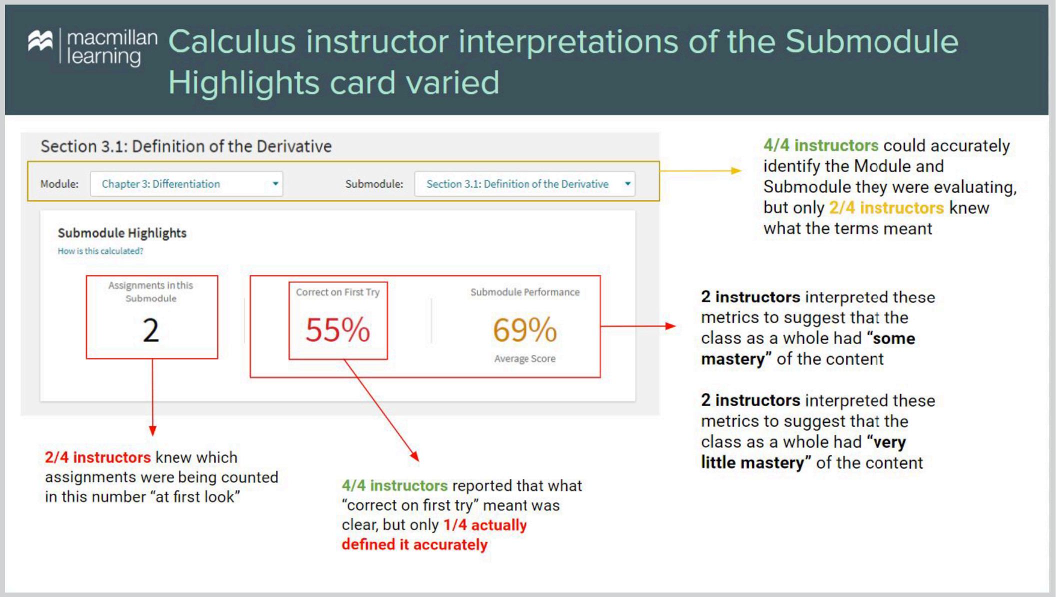
APPENDIX

Table A1. Participating instructor characteristics

	#	%
DIscipline teaching		
Biology	6	16
Calculus	4	11
Chemistry	4	11
Economics	24	63
Years Teaching		
1-5 years	6	16
6-10 years	10	26
11-15 years	5	13
More than 15 years	17	45
Comfort with educational technology		
Extremely uncomfortable	0	0
Uncomfortable	4	11
Comfortable	13	34
Extremely comfortable	21	55
Agreement that publisher provided digital learning tools enhance pedagogy		
Agree	13	34
Strongly agree	25	66
Used a published provide learning tool last time they taught this course?		
No	6	15
Yes	32	84

Note: n=38 instructors

Example insights derived from perception items on the dashboard report


macmillan learning Calculus instructor interpretations of the Submodule Highlights card varied

Section 3.1: Definition of the Derivative

Module: Chapter 3: Differentiation ▾ Submodule: Section 3.1: Definition of the Derivative ▾

Submodule Highlights

How is this calculated?

Assignments in this Submodule	Correct on First Try	Submodule Performance
2	55%	69% Average Score

4/4 instructors could accurately identify the Module and Submodule they were evaluating, but only **2/4 instructors** knew what the terms meant

2 instructors interpreted these metrics to suggest that the class as a whole had **"some mastery"** of the content

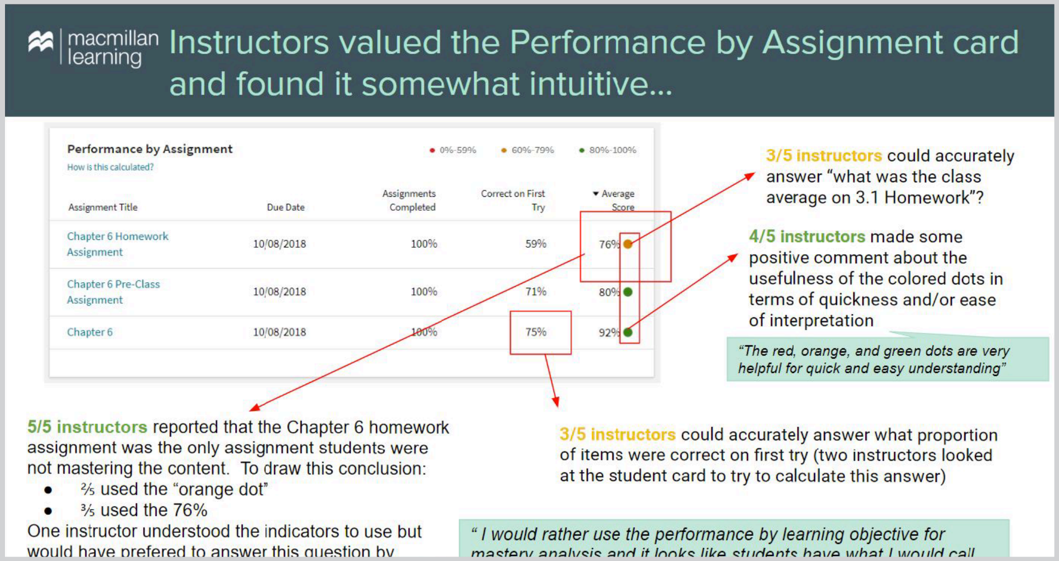
2 instructors interpreted these metrics to suggest that the class as a whole had **"very little mastery"** of the content

2/4 instructors knew which assignments were being counted in this number "at first look"

4/4 instructors reported that what "correct on first try" meant was clear, but only **1/4 actually defined it accurately**

Image 20: Weekly readout of insights derived from the evaluation

Example insights derived from performance items on the dashboard report


macmillan learning Instructors valued the Performance by Assignment card and found it somewhat intuitive...

Performance by Assignment

How is this calculated?

● 0%-59% ● 60%-79% ● 80%-100%

Assignment Title	Due Date	Assignments Completed	Correct on First Try	▾ Average Score
Chapter 6 Homework Assignment	10/08/2018	100%	59%	76% ●
Chapter 6 Pre-Class Assignment	10/08/2018	100%	71%	80% ●
Chapter 6	10/08/2018	100%	75%	92% ●

3/5 instructors could accurately answer "what was the class average on 3.1 Homework"?

4/5 instructors made some positive comment about the usefulness of the colored dots in terms of quickness and/or ease of interpretation

"The red, orange, and green dots are very helpful for quick and easy understanding"

5/5 instructors reported that the Chapter 6 homework assignment was the only assignment students were not mastering the content. To draw this conclusion:

- ⅖ used the "orange dot"
- ⅗ used the 76%

One instructor understood the indicators to use but would have prefered to answer this question by

3/5 instructors could accurately answer what proportion of items were correct on first try (two instructors looked at the student card to try to calculate this answer)

" I would rather use the performance by learning objective for mastery analysis and it looks like students have what I would call

Image 21: Weekly readout of insights derived from the evaluation

About the Authors

Dr. Kara McWilliams
Vice President 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. Billie-Jo Grant

Billie-Jo Grant is a Senior Impact Research Scientist at Macmillan Learning. She has over 15 years of experience in educational research. Her primary areas of interest include educational advocacy, policy research and analysis, research and evaluation methodologies, and program evaluation. Billie-Jo earned her PhD from the University of Virginia in Educational research, statistics, and evaluation. She also holds a Master's degree in Social Foundations of Education and a Bachelor's degree in Economics, both from the University of Virginia. Billie-Jo is also a Statistics Professor at California Polytechnic University.

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