AI-Generated Questions in Context: A Contextualized Investigation Using Platform Data, Student Feedback, and Faculty Observations

Rachel Van Campenhout, Benny G. Johnson, Michelle Clark, Melissa Deininger, Shannon Harper, Kelly Odenweller, and Erin Wilgenbusch

Original scientific article

***Abstract*—In recent years, artificial intelligence has been leveraged to develop an automatic question generation (AQG) system that places formative practice questions alongside textbook content in an ereader platform. Engaging with formative practice while reading is a highly effective learning strategy. AQG made it possible to scale this method to thousands of textbooks and millions of students for free. Previous research studies used aggregated data from all questions answered by all students to complete the largest evaluation of the performance metrics for automatically generated questions. However, these studies also indicated that when assigned in a classroom setting, student behavior and question performance metrics would differ. In this study, we evaluate data collected from 19 course sections taught by four faculty members at Iowa State University to gain a broader understanding of how students engage with these AI-generated practice questions when part of their university courses. Implementation strategies for the courses, student engagement, and question performance metrics are analyzed, and student feedback gathered from surveys and course evaluations are presented. Implications for further use in higher education classrooms are discussed.**

***Keywords*—automatic question generation, performance metrics, question difficulty, persistence, natural learning context, student behavior.**

I. Introduction

A primary component of many higher education classrooms is a textbook, which faculty expect students to read, learn from, and apply their learning to assignments and assessments. However, while textbooks have been the gold standard of learning content, they present some challenges, such as engagement and active learning. First, textbooks are often assigned by instructors yet students do not read them as intended [1, 2, 3, 4]. Data from etextbook platforms confirmed low student reading, and only varying success from traditional instructor strategies—such as reading quizzes or discussions—for increasing engagement [5]. However, assigning formative practice was found to increase engagement over any other reported strategy [6]. Second, as passive learning environments, they are not the most effective way of learning. Research from Carnegie Mellon University's Open Learning Initiative found that incorporating formative practice into text content in a learning by doing approach has been shown to be six times more effective for learning than reading alone [7, 8]. This doer effect learning science principle was further found to be causal to learning [8, 9]. Replicated research on the doer effect confirms that this learning by doing method is generalizable and should be provided to as many learners as possible [10, 11, 12].

However, formative practice incorporated with digital textbook content is not currently common for students in higher education. Learning environments such as courseware are highly effective at delivering a learning by doing experience, but are typically difficult to scale due to the cost of development and barriers to adoption. The advances in artificial intelligence have made it possible to generate the volume of formative practice needed for the learn by doing method. Automatic question generation (AQG) systems have been increasing in popularity for research groups globally for a variety of educational purposes [13]. While there are varying approaches to generation and applications for use, Kurdi et al. 13] noted that no clear gold standard was identified for automatically generated (AG) questions—this contributed greatly to the lack of evaluation of AG questions using student data.

An AQG system was developed that uses the textbook as the content for the generation of matching and fill-in-the-blank (FITB) formative question types (evaluated herein). These AG questions were initially placed in courseware learning environments alongside human-authored questions and evaluated across six courses, finding that there was no difference in how students used the AG and human-authored questions on key performance metrics: engagement, difficulty, persistence, and discrimination [14, 15]. These AG questions were then placed in the Bookshelf ereader alongside etextbook content as a study feature named CoachMe. In the largest analyses of AG questions using student data known to date, prior performance metrics were mirrored, confirming these benchmarks at scale [16]. It is notable that in the initial research comparing AG and human-authored questions [14], as well as subsequent research on CoachMe questions [16, 17], that the largest differences in performance metrics were due to the cognitive process dimension of the question type; there was typically no practical difference based on whether questions

Manuscript received January 27, 2025; revised February 28, 2025. Date of publication April 22, 2025. Date of current version April 22, 2025.

The paper was presented in part at the International Conference on Software, Telecommunications and Computer Networks (*SoftCOM*) 2024.

R. V. Campenhout, B. G. Johnson, and M. Clark are with the Learning Science Department, VitalSource, USA (e-mails: {rachel.vancampenhout, benny.johnson, michelle.clark}@vitalsource.com).

M. Deininger, S. Harper, K. Odenweller, and E. Wilgenbusch are with the Iowa State University, USA (e-mails: {mdein, sharper, kellyo eew}@iastate.edu).

Digital Object Identifier (DOI): 10.24138/jcomss-2024-0120

were AG or human-authored. The matching are a recognition type and typically had higher engagement, difficulty, and persistence rates, while the FITB are a recall type and typically had lower mean rates. While the difference between recognition and recall question types has been researched for decades and the classifications made clear [18, 19], this research simply contributes additional examples of differing question performance and student behaviors. Additional research into student interaction patterns with these questions [17] and feedback [20] revealed new insight into student behavior and learning patterns.

Research using natural learning contexts is valuable for its contributions to external validity and generalizability [7, 12]. Research in the classroom context does not risk altering natural student behaviors as can happen in controlled or semi-controlled experiments, and reduces the ethical concerns of withholding treatments expected to be beneficial for students. While beneficial for identifying performance benchmarks, the research using an aggregated data set from hundreds of thousands of students and millions of answered questions is not necessarily representative of what may occur in a classroom either. These research studies indicated that students in a university course using the CoachMe questions used the questions differently, resulting in higher mean first attempt accuracy and persistence [16] and different interaction patterns [17]. Given the common utilization of etextbooks as the primary learning resource in courses, it is key to understand how the AG questions perform in classroom environments, especially considering how varying course contexts and instructor implementation strategies can greatly impact student engagement and learning [21, 22]. The goal of this paper is to provide a comprehensive understanding of the use of AG questions in the classroom by looking at several types of data: general student engagement with the textbook, question difficulty and persistence, interaction patterns, non-genuine student answers, and faculty observations.

This paper presents an extension of research presented at the 32nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2024) [36], featuring student perception as gathered by surveys and course evaluations. There are several contributions to educational research made by this paper. First, this study utilizes student data from natural contexts—from 19 courses between Fall 2022 and Spring 2024 at Iowa State University—which is not yet common in evaluation studies of AG questions. Second, we expand on what is known about how AG questions perform through platform data analysis that include engagement, difficulty, and persistence. Third, insights into student behaviors are gained by investigating student interaction patterns, thumbing rates, and non-genuine answers. Finally, this study contributes a unique insight into student perceptions of the questions through multiple feedback channels that, combined with quantitative analyses, provides a rare holistic view of AG questions as a study tool.

This paper is organized as follows: the methods are outlined in section II; results are in section III, including engagement, performance metrics, interaction patterns, non-genuine response rates, student ratings, student survey feedback, course evaluations, and faculty observations; discussion and conclusion are in section IV.

II. Methods

The AQG system used in this paper is an expert-designed, rule-based system. Neither question type evaluated in this paper was generated using large language models. The course textbook is used as the corpus for natural language processing. The system uses both syntactic and semantic information to identify important sentences and key terms, then a rule system is applied to transform these into questions (for additional details on the AQG system, see [14, 16]). Once the questions are generated, they are placed alongside the corresponding section of the textbook. As seen in Figure 1, students receive immediate

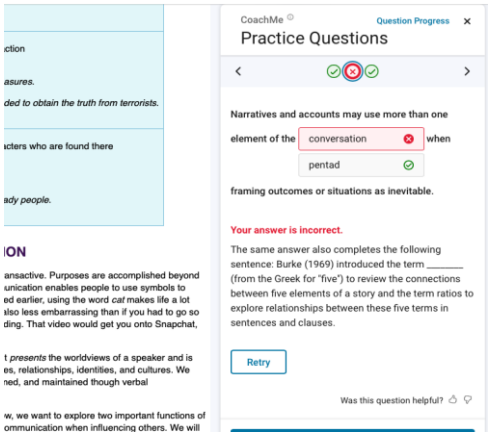


Fig. 1. Question example showing immediate, scaffolded feedback. Options for retrying, revealing the answer, and rating the question are also available in this panel.

feedback after they submit an answer. For FITB questions, scaffolding feedback that uses an additional content example from the textbook has been shown to perform best for student persistence and increasing second attempt correct response rates [20]. After an incorrect response, students can either retry the question, which resets the question, or reveal the answer (with an additional retry optional). Students are able to monitor their progress on a progress panel that shows the percentage of practice completed, correctness states for each question, and allows for navigation to different questions and question sets. The progress panel has been shown to help increase student motivation to complete more or all of the available practice that was required [23].

As students read the etextbook and answer questions, the ereader platform is collecting clickstream data, attaching a timestamp to each action students take. This contextual microlevel data is invaluable for educational data science [24, 25], helping to answer new educational questions or old questions in new ways [26]. In this paper we aim to use this data for both old and new questions: the long-considered issue of textbook engagement and benefits of formative practice and the emerging field and research on automatic question generation.

In order to better understand the impact of automatically generated formative practice in university classrooms, the VitalSource learning science team partnered with Iowa State University faculty. The goal of this partnership was to learn about successful implementation practices across varied course contexts, gather student faculty perceptions on the practice questions as a learning tool, and gain insight into its benefits for learning in general. The partnership began with pilot courses in

TABLE I. COURSE LEVEL DETAILS

Instructor	*Course Title*	*Textbook*	*Modality*	*Course Selection*	*Credit*
Melissa Deininger	INTST 235-Introduction to International Studies	Crossing Borders: International Studies for the 21st Century [27]	Online Asynchronous	Mix of required for majors and elective for non-majors	10 pts/chapter for 80% completion
Shannon Harper	CJ 406-Gender and Crime	Women, Gender, and Crime: Core Concepts [28]	In person	Special topic choice for majors	10 pts/chapter for 80% completion
Kelly Odenweller	HDFS 270-Family Communication and Relationships	Family Communication: Cohesion and Change [29]	Hybrid Synchronous	Elective for majors	1-10 points for completing 10-100% of practice per chapter
Kelly Odenweller	COMST 101-Introduction to Communication Studies	Communication in Everyday Life: A Survey of Communication [30]	Hybrid Synchronous	Required for majors	1-10 points for completing 10-100% of practice per chapter
Erin Wilgenbusch	PR 321-Public Relations Writing	Becoming a Public Relations Writer: Strategic Writing for Emerging and Established Media [31]	In person	Required for majors	Extra credit
Erin Wilgenbusch	PR 424-Public Relations Campaigns	Strategic Planning for Public Relations [32]	In person	Capstone course	Extra credit

the summer of 2022 (including two authors of this paper) and continued each semester from the fall of 2022 to spring of 2024 (including all four faculty authors). Each faculty member met with the senior research scientist at VitalSource to review the feature and discuss each course and implementation approach. Faculty were sent custom data reports weekly that gave the percentage of practice completed by each student, which was used to assign points (completion only). At the end of the term, students optionally provided feedback to a survey created by the VitalSource team, and the faculty met once more with the research scientist to review the semester and provide feedback.

This study includes four faculty members who taught 19 undergraduate course sections of six courses, for a total of 2,090 students between the fall of 2022 and spring of 2024. The context varies greatly between courses, with key details captured in Table I. The courses were in varying subject domains, course classifications, delivery modalities, and had different strategies for assigning practice completion credit. For example, Prof. Harper taught traditional face-to-face 16-week sections of a special selection choice for majors, while Prof. Deininger taught large asynchronous 8- and 16-week concurrent sections of a course that could be used to fulfill a university requirement as well as a selection for majors. Prof. Odenweller taught large courses that were elective or required for majors while Prof. Wilgenbusch taught a small capstone course for seniors and a required writing course for majors. As these differences are common to all universities, the implementation of these AG questions as an assigned part of these courses helped to reveal trends that would be useful for generalization purposes.

III. RESULTS

A. Engagement

Prior to more detailed questions about the performance of AG questions is a much simpler question: how did assigning the AG formative practice impact student engagement with the textbook? There were several instances of courses where the same etextbook was used without the questions in a prior semester. For three courses with a reasonable prior semester comparison, the semester with the questions assigned had an average of three times the days used over prior terms with no questions. Figure 2 shows an example of a COMST 101 course where the mean days used went from five in a prior course to 20 in the course with questions assigned.

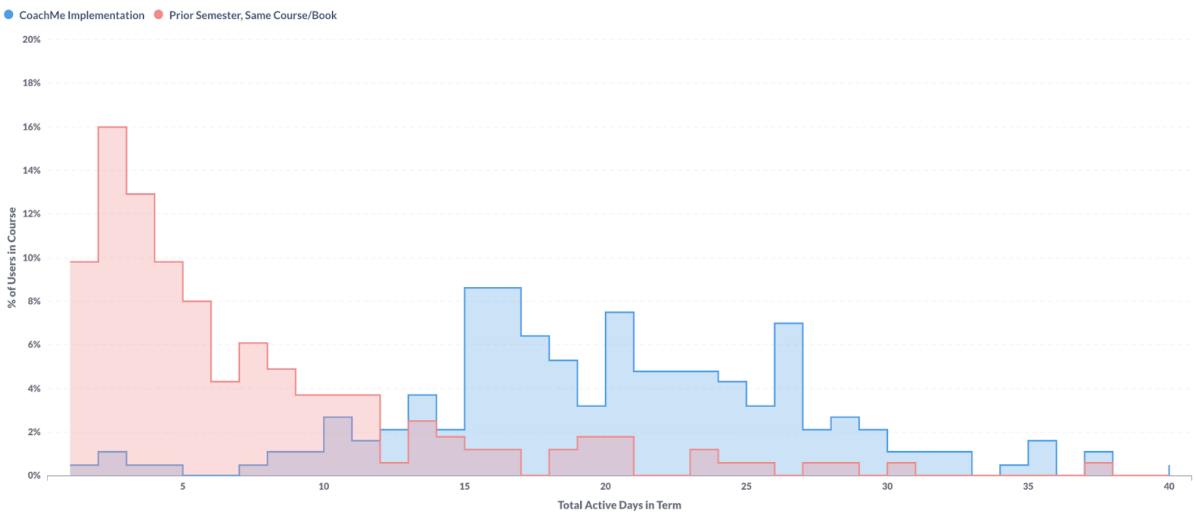


Fig. 2. A histogram comparison of textbook days used for a prior semester of COMST 101 without the AG questions available (red) to the semester that had them assigned (blue).

B. Performance Metrics

A primary research area for AG questions is their performance as learning objects for students. This can be investigated by looking at key performance metrics such as difficulty and persistence. In this analysis, we refer to the difficulty index, which is the percentage of students who answered a question correctly (the higher the difficulty index value, the easier the question). Persistence is a subset of the difficulty data that looks at when students get questions incorrect on their first attempt, how often do they continue to answer until they get the correct response. These two metrics combined not only give benchmark data for AG question performance, but insight into student behavior.

A prior analysis of all student-question interaction data points collected between January 2022 and April 2023 looked at a total of more than 329,000 students, 901,000 individual questions, and 8.4 million total question events in over 8,000 textbooks to evaluate question performance at scale [16]. In that study, the matching questions had a mean difficulty of 79.3% and a persistence of 69.5% while the FITB questions had a mean difficulty of 54.7% and persistence of 58.5%. These values were hypothesized to not be entirely reflective of a classroom implementation, however, as they included students in any learning context, with engagement mostly optional, not required. The performance metrics of the 19 courses included here confirm that hypothesis. Across all courses, the matching questions had a mean difficulty of 82.8% and a persistence of 96.7% and the FITB questions had a mean difficulty of 82.7% and a persistence of 94.0%. These values are much higher than the aggregated data, suggesting that when assigned—even only for completion—that students more seriously attempt the questions and persist until entering the correct response.

Table II shows a detailed breakdown of these performance metrics by course. Despite the differences in course, semester, class size, and number of questions, the mean difficulty for matching and FITB is consistently in the upper 70 to low 90 percent range. Interestingly, there are some courses where the matching had a lower difficulty index—meaning more difficult—than the FITB. This is contradictory to the trend from aggregated data and worth further investigation.

Persistence (the rate at which students who incorrectly answer continue on to input a correct response) was generally high across courses, and typically similar between matching and FITB. There are instances where both are similarly high as in INTST 235 F23-7, or both similarly low as in HDFS 270 S23, indicating persistence rates are at least partially related to that cohort of students. Matching questions had a perfect 100% persistence for three courses, and ten over 97%. FITB questions had nine courses over 97%. These incredibly high persistence rates are across widely varied course sizes and contexts.

TABLE II. PERFORMANCE METRICS BY COURSE

Section	*Students*	*Question Total*	*Total Answered*	*Matching Mean*	*Matching Persistence*	*FITB Mean*	*FITB Persistence*	*FITB Non-Genuine Answers*	*FITB Non-Genuine Persistence*
PR 424 F22	32	231	6247	80.22	89.9	87.9	92.9	8.7	93.3
PR 321 F23	13	183	1057	79.0	100.0	90.1	91.1	8.9	100.0
PR 321 S24	11	312	1441	80.8	100.0	74.5	98.4	48.3	100.0
COMST 101 S23	181	183	27444	72.7	96.3	80.2	95.0	32.2	99.5
COMST 101 F23	146	165	19864	73.3	92.9	72.9	88.8	34.2	92.6
COMST 101 S24	181	243	37184	77.8	98.6	82.8	99.1	32.7	99.7
HDFS 270 F22	49	266	2792	88.2	95.1	76.4	82.6	14.1	97.5
HDFS 270 S23	59	266	14553	84.0	89.2	85.2	74.0	34.5	78.4
HDFS 270 F23	68	254	16577	82.8	95.9	84.9	95.3	38.2	99.2
HDFS 270 S24	61	371	19513	85.1	91.4	79.3	91.2	38.1	88.3
INTST 235 S23-5	342	202	59813	80.8	98.3	78.0	96.2	41.3	97.2
INTST 235 F23-7	68	185	10919	77.5	99.0	85.2	99.3	40.6	99.2
INTST 235 F23-4	325	185	51054	76.1	97.0	77.7	94.6	33.0	99.0
INTST 235 S24-6	191	300	47262	84.8	99.6	84.5	99.0	53.5	100.0
INTST 235 S24-5	183	300	44703	87.9	97.3	88.0	98.5	49.9	100.0
CJ 406 F22	70	165	6282	91.7	100.0	81.7	97.6	17.4	100.0
CJ 406 S23	39	198	7018	87.9	99.5	82.7	98.0	13.0	99.1
CJ 406 F23	50	199	8685	90.6	98.4	86.1	97.8	9.8	100.0
CJ 406 S24	42	228	8817	92.1	98.8	92.9	97.5	3.6	95.0

C. Interaction Patterns

The data can also reveal new insights into student behavior by investigating interaction patterns. When a student inputs an incorrect response, there are several options for the next action. As seen in Figure 1, the student could retry the question on their own, reveal the answer, reveal the answer then retry, a combination of the above, or simply abandon the question without further action. In prior research on an aggregated data set, the most popular interaction pattern after an incorrect response was to reveal the answer with no further action [17]. When we evaluate the top two patterns for these 19 courses they are [incorrect → reveal → retry correct] and [incorrect → retry correct] for both question types. For the FITB questions, the top pattern was always [incorrect → reveal → retry correct] at 60.2% of all initially incorrect interaction patterns. However, the matching questions were split with 11 courses sharing the top pattern of [incorrect → reveal → retry correct] but the remaining eight courses having [incorrect → retry correct] instead. This indicates students tend to be more willing to retry matching on their own. In general, FITB has a much higher immediate reveal rate and much lower rate of completing correctly without a reveal, likely because it is a recall type and matching is recognition type. However, what is notable about these top interaction patterns is that they end in the students inputting the correct response as the final action. This is contrary to what was noticed in the aggregated study, suggesting students in the classroom have a motivation to persist to complete the question to a correct state.

D. Non-Genuine Responses

The non-genuine response rates are an area of investigation that reveals insight into student behavior. When assigning points—even for completion—it becomes a valid concern of whether students are genuinely attempting the questions or just "gaming the system" to get their points. In order to better understand student behavior, we investigate the questions students get incorrect on the first try. In all these courses, that percentage is relatively low to begin with (17.2% and 17.3% incorrect for matching and FITB respectively). Of those incorrect responses for FITB, we analyzed the text inputs based on a few simple rules such as answers under three characters, no vowels, punctuation, and known responses such as "idk". While not perfect, these rules identify the vast majority of what we consider non-genuine responses, or students inputting responses they know are not going to be correct.

The non-genuine response percentages vary widely across the 19 courses in Table II. Four courses have rates under 10%, while five courses have rates over 40%. Why there are such varying rates is unknown; it could be related to the course context, the questions generated for the textbook, or the nature of the students in the course. Recalling the non-genuine response rate is already taken from the subset of incorrect responses (a low percentage to begin with), the non-genuine responses still comprise the small minority of all responses even when the percentages are over 40. But what is more interesting is the persistence rate for this non-genuine response rate. The mean persistence rate for the non-genuine responses is 96.7%, with several courses boasting 100%. This means that when students input a non-genuine response, they almost always continued to work the question to input the correct response. Recalling the interaction pattern data, this included both cases where students revealed the answer then retried, or retried on their own to get the correct response. Why, then, do students input non-genuine responses? There are certainly many reasons that include students who may be rushing through without taking the time to think of the response on their own. However, it also seems to be a way to see feedback or reveal the answer before typing it in themselves—both of which are still contributory to the learning process.

E. Student Ratings

After students have made a first attempt to answer a question, a rating option is presented beneath the question, as seen in Figure 3. Students can select a simple thumbs up or down icon to rate the question. If they rate a thumbs down, they can provide optional follow up feedback.

The student ratings have several practical uses. First, the student ratings are used for iteratively improving the question set in the textbook. An adaptive platform-level system was developed to monitor all questions in all textbooks in real time (something that could not be achieved through human efforts). This content improvement system uses student rating data to determine if questions should be removed and replaced [33]. The overall thumbs down rate from 3,594,408 questions answered is just 0.194%% [33]. Secondly, this rating data is analyzed in a regression model to determine features of questions that cause students to rate up or down, which provides practical guidance for improving the AQG system [34].

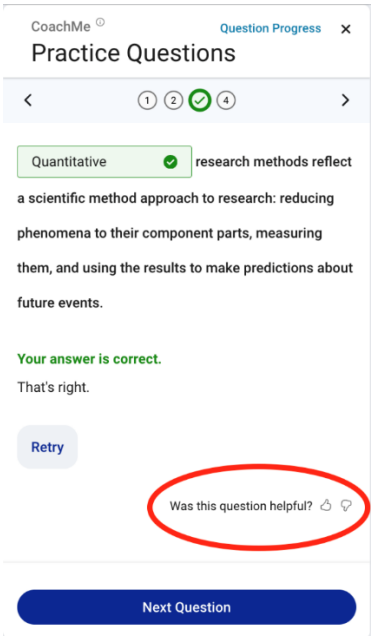


Fig. 3. An answered question with the rating question circled.

The overall rating of questions across all 19 sections was similarly small. There were 18 thumbs up ratings and 60 thumbs down ratings. There were 219,463 opportunities for students to rate questions, which gives the thumbs up rate 0.08/1000 and the thumbs down rate of 0.27/1000. These rates are both much lower than the aggregated thumbing rates, and also the higher rate is opposite what is typically seen [33].

F. Student Survey Feedback

At the end of each semester, faculty were given a product survey to send to students. The survey was optional and completely anonymous. Delivery methods varied with some faculty giving the survey during class and others sending it via email. The goal of the survey was to gather student perceptions of the questions as a learning tool. Of the 19 course sections, three sections did not respond to the survey, resulting in a final group of 16 courses with a total of 744 student responses. Not all survey question results will be reported for brevity.

The survey began with a few questions on the students' general perceptions to provide context useful for interpreting later feedback on the formative practice. This includes questions on print versus etext preferences, difficulty of the course, and anticipated grade. These questions were useful for interpreting responses about the formative practice. One of the first questions asked students, "Generally, how do you like using digital textbooks as a learning resource compared to print books?" provided initial insight into student preferences. While the responses in each course varied somewhat, the aggregated totals revealed 39.4% preferred etextbooks much better, 30.7% somewhat better, 17.5% about the same as print, 10.1% somewhat less than print, and 2.2% much less than print. Understanding that a small group of students in most classes preferred the print modality was a useful lens, as we found later in the survey that same percentage of students typically did not think practice was beneficial. Students may prefer print for varying reasons and that attitude colors the feedback on other questions.

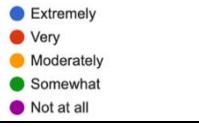
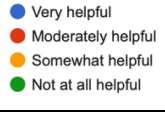
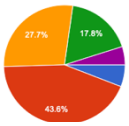
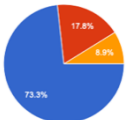
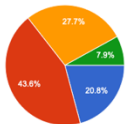
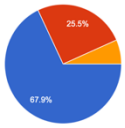
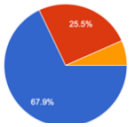
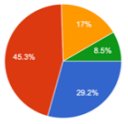
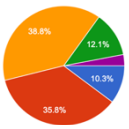
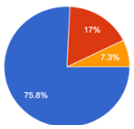
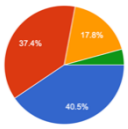
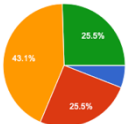
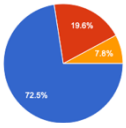
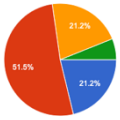
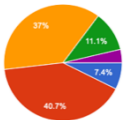
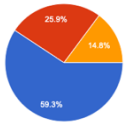
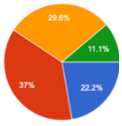
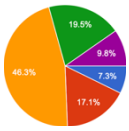
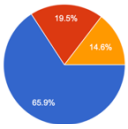
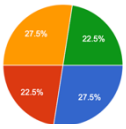
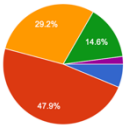
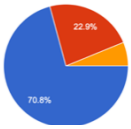
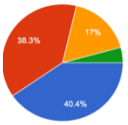
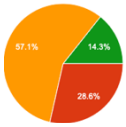
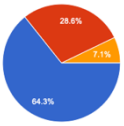
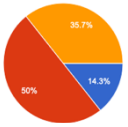
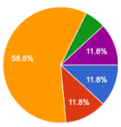
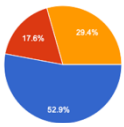
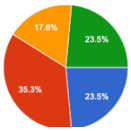
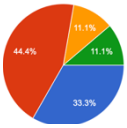
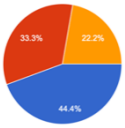
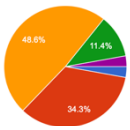
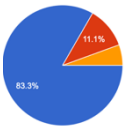
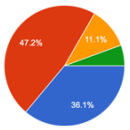
Students were asked, "How important do you think reading the textbook is to your overall success in this course?" This question aimed to interpret not their perception of the textbook itself, but their conceptualization of the relationship between the textbook and their success. This could capture how students felt the textbook was incorporated into their course or how students felt about their own abilities to succeed with or without the textbook. Responses varied widely between courses—likely due to differences between courses as well as the student cohort. The aggregated results revealed 11% thought the textbook was extremely important, 35.3 thought it was very important, 36.3% thought it was moderately important, 14.3% thought it was somewhat important, and 3.1% thought it was not at all important to their success. Table III includes visual responses for this question and others for the COMST 101, HDFS 270, and CJ 406 courses.

Next, students were asked, "In general, do you think doing practice questions while reading is helpful for learning?" This question was included to get a general sense of students' perception of learning by doing, not the questions provided. Across courses, 72% responded yes, 19% responded maybe, and 9% responded no. The majority of students believing in this learning by doing approach is great, as the doer effect has proven its benefit for learning and increasing learning outcomes. The survey results for students who responded maybe or no raises the question, is there more we should be doing to educate students on the benefits of this method? [35] found students overestimate the explicit learning value of reading and underestimate active learning (formative practice). What is very interesting about these results is the remarkable consistency across courses. As seen in Table III, the column with pie charts for this question show a similarity in the proportion of responses. This is more interesting given the wide variability in responses between courses for how important students thought the textbook was to their success. No matter how a class responded to the textbook importance question, their views on doing formative practice being good for learning was consistent.

The next question also has pie chart visuals reported in Table III: "How helpful did you find the practice questions for studying and preparing for assignments?" The aggregated results found 32.4% of students selected very helpful, 39.5% selected moderately helpful, 20.5% selected somewhat helpful, and 7.5% selected not at all helpful. Note the percentage of students who didn't find the practice helpful is similar to those who thought practice wasn't beneficial for learning. When comparing the pie charts for this question to how beneficial students view practice in general, there are some consistent similarities. The proportion of students who reported the practice as very and moderately helpful tends to be very close to the proportion of students who generally think practice is helpful. The proportion of students who felt the practice was somewhat helpful is similar to those who thought maybe practice was beneficial. The same is seen for students who did not think practice is beneficial and who did not find the formative practice helpful for learning. The similarities between these question responses are positive; it could be conceivable that students perceive practice as beneficial for learning, yet not find these practice questions helpful. To see alignment between general beliefs and perceptions on the helpfulness of the AG questions is a validation for their use.

We asked students to think about their future use of the feature: "If the CoachMe practice questions were available in a textbook for a future course you take, how likely would you be to use them, even if not assigned?" In total, 15.6% said extremely likely, 25.6% said very likely, 41.5% said moderately likely, and 17.4% said not at all likely. The overall proportions show similarities to student perceptions on digital versus print books and perceptions on the benefit of practice. While experience has taught us students typically do not do things that are not assigned, that student sentiment for the practice produces positive intention is a measure of success for student perception.

TABLE III. STUDENT SURVEY RESPONSE VISUALIZATIONS ACROSS COURSES

Section	How important do you think reading the textbook is to your overall success in this course?	In general, do you think doing practice questions while reading is helpful for learning?	How helpful did you find the practice questions for studying and preparing for assignments?
			
COMST 101 S23			
COMST 101 F23			
COMST 101 S24			
HDFS 270 F22			
HDFS 270 S23			
HDFS 270 F23			
HDFS 270 S24			
CJ 406 F22			
CJ 406 S23			
CJ 406 F23			
CJ 406 S24			

G. Course Evaluations

Students also offered feedback on the formative practice through other avenues. At the end of each course, students are able to anonymously submit a course evaluation. The evaluations are reviewed by the department chair before being made available to the instructor. Student feedback at this level is key to understand what did and did not work for students. In several courses, students commented on the questions as homework.

What is helping me learn in this class? (HDFS 270)

- …Doing the Coach Me questions is also helpful to see the information we discussed in lecture in a different format.
- I think the coach me questions help a lot because I feel like without the coach me questions I wouldn't read the textbook and I would just rely on the lecture slideshow.
- I really like the CoachMe questions on each chapter. I feel like that helps me find and remember keywords for the quizzes and assignments.
- The coach me questions and lecture both helping me understand the material.
- In this class, the CoachMe questions have really helped me learn and apply the information….
- …I also think the CoachMe questions are holding me accountable for reading the textbook because without them l probably wouldn't make it a priority….
- The textbook and coach me questions as well have helped me learn more about the subject we are learning about in lecture.

What is helping me learn in this class? (INTST 235)

- The book questions holding you accountable to read the text
- I feel it is difficult to truly learn material for an online course, but having assigned quizzes as well as book questions for our online book has helped me learn most of our content so far for this class.
- The book questions helped me learn content the best in this class as it helped me to pay more attention to the book and comprehend the information.
- The book questions helped me learn the most.
- The book questions were honestly very helpful to spark ideas when it came to discussion posts.
- Being that this class was 100% online, the book questions and quizzes helped me test my knowledge.

How am I contributing to my learning? (HDFS 270)

- After class I always go back to my dorm and write the notes we went over during the lecture. I also read the chapter that we are discussing during the week and complete my CoachMe questions and quizzes. I also make it a priority to be in class every time we meet so I can learn the material to the best of my ability.
- I am contributing to this class by showing up to every single class and doing my weekly work outside of class. I read the textbook, fill out the CoachMe questions, write my papers, and take my quizzes with the extra time we have outside of class.
- I have been reading my textbook and studying for my quizzes, which has helped me a lot in the course so far. I have failed to do that in the past with some courses and have really seen a difference in this semester.

What do I need to do to improve my learning in this course? (HDFS 270)

- One thing I could do to improve my learning in this course is to read the textbook more often. While I do read parts of it when I do Coach Me questions, I don't read much of it and know it would help retention to see the content in this format.

H. Faculty Observations

The quantitative data collected by the platform reveals important findings around student engagement with the textbook and automatically generated questions. However, in a classroom context, it is also valuable to understand the impact on student learning from a qualitative perspective. Faculty observations on student perceptions and behaviors provide data no other source can contribute. First, faculty agree that simply having data insights into student use of the textbook and a method of holding students accountable for the readings was a significant benefit. Faculty noticed positive changes in a range of ways: students cited the textbook frequently in discussion posts (not something done often in previous courses), class discussions were more engaging, and students' first drafts on written projects required less feedback and major changes than prior courses. In COMST 101, student reading quiz scores increased compared to a prior semester. These observed benefits to student engagement in other areas of the classroom show the value in assigning the AG questions as a learning by doing tool that both engages students and holds them accountable for using their primary learning material as faculty intended.

Faculty and student feedback was overall positive, but not exclusively. The questions generated initially for PR 424 included a higher proportion of questions related to examples in the textbook which students did not find as helpful. As a research partner, this feedback was given directly to the learning science team and resulted in updates to the generation process to improve the question set. Automatically generated questions cannot be perfect (nor can human-authored questions) so iterative improvement is critical for continually optimizing the question sets [33]. In a course evaluation, one student commented, "The only thing that I see could be improved are the CoachMe questions. I think they are great because like I said before they hold me accountable for doing the reading, but I wish the questions were based more around the general concept of the idea we are learning instead of just different phrases from the text. I think this would allow me to better understand the material." This student was commenting on the fill-in-the-blank questions and expressing a want for broader, higher-order questions. This feedback is beneficial to motivate new question development in the future. Students also gave other usability feedback around discoverability which was also used to make updates to the interface and onboarding experience. Collaborative partnerships between education technology developers and faculty and students is key to making improvements to learning tools over time—ultimately benefiting learners.

IV. Discussion and Conclusion

The research using an aggregated data set that included hundreds of thousands of students and millions of answered questions provided important benchmarks for AG question performance as well as interesting insights into student behavior [16]. Yet it is equally valuable to investigate the performance of these questions in natural learning contexts and the benefit this learning by doing feature has for students and faculty. Across the 19 courses in this study, there are notable differences in performance metrics compared to the aggregated data of prior research. Both the matching and FITB questions had greatly increased difficulty index boundaries; students answered correctly much more frequently in each of these courses than in the aggregated data set. Students also persisted to answer correctly nearly every time they were incorrect—another difference in behavior from the aggregated data. The classroom context certainly changes the performance metrics.

When students answer incorrectly, we also see different behaviors. The top two interaction patterns for matching and FITB in a classroom environment both ended in a correct response being entered, which was not the result from the aggregated data [16]. The mean percentage of non-genuine responses was higher than the aggregated data set (29.1% versus 12.2%) [16], but the range between the courses here was 3.6% to 53.5%—a sizable difference. The reason for this range is not known, but it shows the difference that course context can make. What was universal, however, was that when students used a non-genuine answer strategy, they always completed the question correctly afterward.

Student perceptions of the AG questions as a learning tool are equally important to consider in addition to performance metrics. The survey data revealed interesting insights into student preferences and perceptions. A small group of students prefer ebooks less than print books, and this same small proportion of students tended to not think learning by doing was beneficial for learning and did not find the questions helpful. Students across courses had varying opinions on how important the textbook was to their overall success in the course, however, most students (an average of 72%) thought that generally, doing practice was beneficial for learning. This perception on learning by doing was surprisingly consistent across courses—a stark contrast to the variation in their thoughts on textbook importance. Another positive finding was that nearly the same proportion of students who thought formative practice was generally good also found the AG formative practice helpful for studying and learning. Students gave their feedback voluntarily through other channels like course evaluations. A formal way for the university to solicit feedback, students reported that doing the practice held them accountable for doing the reading and helped their comprehension of the textbook content.

Classrooms are complex learning environments with an incredible number of variables that impact student behavior and outcomes—precisely why research in natural learning contexts is important to understanding the impact of learning tools. In most higher education courses, the textbook is intended to be the primary learning resource and yet research has validated that students often do not read the textbook. By adding automatically generated practice to etextbooks, students had access to a learning by doing method to stay actively engaged with the learning material while reading, and faculty had a way to hold students accountable for completing the assigned reading. Assigning points for completion of the automatically generated practice places value for the student on doing the reading and practice, which is critical for their learning. In these courses, we found increased usage of the textbook, difficulty and persistence metrics indicating students were genuine in their attempts at the practice, and observational indications that students were benefiting from this assigned practice in other areas of the classroom experience. This research contributes to the literature that formative practice combined with the primary reading content benefits students and faculty alike.

Acknowledgements

The many semesters of this research project were supported by members of the Iowa State University Bookstore—Heather Dean, Emma White, and John Wierson—and we thank them for their dedication to students and their learning. We also thank the Center for Excellence in Learning and Teaching for sharing this ongoing work with faculty across the university.

References

[1] C. M. Burchfield and J. Sappington. 2000. Compliance with required reading assignments. Teaching of Psychology 27, 1 (2000), 58. https://psycnet.apa.org/record/2000-07173-017.

[2] P. A. Connor-Greene. 2000. Assessing and promoting student learning: Blurring the line between teaching and testing. Teaching of Psychology 27, 2 (2000), pp. 84–88. https://doi.org/10.1207/S15328023TOP2702_01.

[3] A. Schneider. 2001. Can plot improve pedagogy? Novel textbooks give it a try. Chronicle of Higher Education 47, 35 (2001), A12.

[4] T. Berry, L. Cook, N. Hill, and K. Stevens. 2010. An exploratory analysis of textbook usage and study habits: Misperceptions and barriers to success. College Teaching 59, 1 (2010), pp. 31–39. https://doi.org/10.1080/87567555.2010.509376.

[5] J.-E. Russell, A. M. Smith, S. George, and B. Damman. 2023. Instructional strategies and student eTextbook reading. In ACM International Conference Proceeding Series, pp. 613–618. https://doi.org/10.1145/3576050.3576086.

[6] N. Brown, R. Van Campenhout, M. Clark, and B. G. Johnson, "Are Students Reading? How Formative Practice Impacts Student Reading Behaviors in Etextbooks," in *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S'24)*, pp. 383–387. https://doi.org/10.1145/3657604.3664668

[7] K. Koedinger, J. Kim, J. Jia, E. McLaughlin and N. Bier, "Learning is not a spectator sport: Doing is better than watching for learning from a MOOC," Proceedings of the Second ACM Conference on Learning@Scale, Vancouver, BC, Canada, 2015, http://dx.doi.org/10.1145/2724660.2724681.

[8] K. Koedinger, E. McLaughlin, J. Jia and N. Bier, "Is the doer effect a causal relationship? How can we tell and why it's important," Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, Edinburgh, United Kingdom, 2016, http://dx.doi.org/10.1145/2883851.2883957.

[9] K. R. Koedinger, R. Scheines and P. Schaldenbrand, "Is the doer effect robust across multiple data sets?" Proceedings of the 11th International Conference on Educational Data Mining, 2018, http://dx.doi.org/10.1145/2883851.2883957.

[10] R. Van Campenhout, B. G. Johnson and J. A. Olsen, "The doer effect: Replicating findings that doing causes learning," Proceedings of eLmL 2021: The Thirteenth International Conference on Mobile, Hybrid, and On-line Learning, 2021. https://www.thinkmind.org/index.php?view=article&articleid=elml_2021_1_10_58001.

[11] R. Van Campenhout, B. G. Johnson and J. A. Olsen, "The doer effect: Replication and comparison of correlational and causal analyses of learning," International Journal on Advances in Systems and Measurements, vol. 15, no. 1&2, pp. 48-59, 2022.

[12] R. Van Campenhout, B. Jerome and B. G. Johnson, "The Doer Effect at Scale: Investigating Correlation and Causation Across Seven Courses," in LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), 2023, https://doi.org/10.1145/3576050.3576103.

[13] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A Systematic Review of Automatic Question Generation for Educational Purposes," International Journal of Artificial Intelligence in Education, vol. 30, no. 1, pp. 121-204, 2020, https://doi.org/10.1007/s40593-019-00186-y.

[14] R. Van Campenhout, J. S. Dittel, B. Jerome, and B. G. Johnson, "Transforming textbooks into learning by doing environments: An evaluation of textbook-based automatic question generation," in Proceedings of the Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education, 2021, pp. 47-56, CEUR Workshop Proceedings, http://ceur-ws.org/Vol-2895/paper06.pdf.

[15] B. G. Johnson, J. S. Dittel, R. Van Campenhout, and B. Jerome, "Discrimination of automatically generated questions used as formative practice," in Proceedings of the Ninth ACM Conference on Learning@Scale, 2022, pp. 325-329, https://doi.org/10.1145/3491140.3528323.

[16] R. Van Campenhout, M. Clark, B. Jerome, J. S. Dittel, and B. G. Johnson, "Advancing intelligent textbooks with automatically generated practice: A large-scale analysis of student data," in *5th Workshop on Intelligent Textbooks. The 24th International Conference on Artificial Intelligence in Education*, 2023, pp. 15-28. [Online]. Available: https://intextbooks.science.uu.nl/workshop2023/files/itb23_s1p2.pdf.

[17] R. Van Campenhout, M. Clark, J. S. Dittel, N. Brown, R. Benton, and B. G. Johnson, "Exploring student persistence with automatically generated practice using interaction patterns," in *2023 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2023, pp. 1-6. doi: 10.23919/SoftCOM58365.2023.10271578.

[18] D. M. Andrew and C. Bird, "A comparison of two new-type questions: recall and recognition," Journal of Educational Psychology, vol. 29, no. 3, pp. 175-193, 1938, https://doi.org/10.1037/h0062394.

[19] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock, "A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives (Complete edition)," New York: Longman, 2001.

[20] R. Van Campenhout, M. Kimball, M. Clark, J. S. Dittel, B. Jerome, and B. G. Johnson, "An Investigation of Automatically Generated Feedback on Student Behavior and Learning," in *LAK '24: Proceedings of the 14th Learning Analytics and Knowledge Conference*, 2024, pp. 850-856. doi: 10.1145/3636555.3636901.

[21] A. Kessler, M. Boston, and M. K. Stein. Exploring how teachers support students' mathematical learning in computer- directed learning environments. Information and Learning Science, 52–78, 121(1–2), 2019, https://doi.org/10.1108/ILS-07-2019-0075

[22] R. Van Campenhout and M. Kimball. At the intersection of technology and teaching: The critical role of educators in implementing technology solutions. IICE 2021: The 6th IAFOR International Conference on Education – Hawaii 2021 Official Conference Proceedings. ISSN 2189-1036, 2021, 151–161. https://doi.org/10.22492/issn.2189-1036.2021.11

[23] R. Van Campenhout, M. Selinger, and B. Jerome, "Designing a Student Progress Panel for Formative Practice: A Learning Engineering Process," in Proceedings of the Third Annual Meeting of the International Society of the Learning Sciences, 2023.

[24] C. Fischer, Z. A. Pardos, R. S. Baker, J. J. Williams, P. Smyth, R. Yu, S. Slater, R. Baker and M. Warschauer, "Mining big data in education: affordances and challenges," Review of Research in Education, vol. 44, no. 1, pp. 130-160, 2020, https://doi.org/10.3102/0091732X20903304.

[25] R. Van Campenhout, B. Jerome, and B. G. Johnson, "Engaging in student-centered educational data science through learning engineering," in *Educational Data Science: Essentials, Approaches, and Tendencies*, A. Peña-Ayala, Ed., Big Data Management. Singapore: Springer, 2023, pp. 1-40. doi: 10.1007/978-981-99-0026-8_1.

[26] D. A. McFarland, S. Khanna, B. W. Domingue, and Z. A. Pardos, "Education data science: past, present, future," AERA Open, vol. 7, no. 1, pp. 1-12, 2021, https://doi.org/10.1177/23328584211052055.

[27] H. I. Chernotsky and H. H. Hobbs, *Crossing Borders: International Studies for the 21st Century*, 4th ed. Thousand Oaks, CA: Sage Publications Inc., 2022.

[28] S. L. Mallicoat, *Women, Gender, and Crime: Core Concepts*, 2nd ed. Thousand Oaks, CA: Sage Publications Inc., 2023.

[29] K. M. Galvin, D. O. Braithwaite, P. Schrodt, and C. L. Bylund, *Family Communication: Cohesion and Change*, 10th ed. New York, NY: Routledge, an imprint of Taylor & Francis, 2019.

[30] S. Duck and D. T. McMahan, *Communication in Everyday Life: A Survey of Communication*, 4th ed. Thousand Oaks, CA: Sage Publications Inc., 2021

[31] R. D. Smith, *Becoming a Public Relations Writer: Strategic Writing for Emerging and Established Media*, 6th ed. New York, NY: Routledge, an imprint of Taylor & Francis, 2020.

[32] R. D. Smith, *Strategic Planning for Public Relations*, 6th ed. New York, NY: Routledge, an imprint of Taylor & Francis, 2021.

[33] B. Jerome, R. Van Campenhout, J. S. Dittel, R. Benton, and B. G. Johnson, "Iterative improvement of automatically generated practice with the Content Improvement Service," in *Adaptive Instructional Systems. HCII 2023. Lecture Notes in Computer Science*, R. Sottilare and J. Schwarz, Eds., Cham: Springer, 2023, pp. 312-324. doi: 10.1007/978-3-031-34735-1_22.

[34] B. G. Johnson, J. Dittel, and R. Van Campenhout, "Investigating student ratings with features of automatically generated questions: A large-scale analysis using data from natural learning contexts," in *Proceedings of the 17th International Conference on Educational Data Mining (EDM)*, 2024, pp. 194–202. https://doi.org/10.5281/zenodo.12729796

[35] P. F. Carvalho, E. A. McLaughlin, & K. R. Koedinger. Is there an explicit learning bias? Students beliefs, behaviors and learning outcomes. Proceedings of the 39th Annual Conference of the Cognitive Science Society (Eds. Gunzelmann, G. et al.), pp. 204–209. 2017. Retrieved from https://escholarship.org/uc/item/00w8g6df

[36] R. Van Campenhout, M. Clark, B. G. Johnson, M. Deininger, S. Harper, K. Odenweller, & E. Wilgenbusch. *Automatically Generated Practice in the Classroom: Exploring Performance and Impact Across Courses.* The *32nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2024)*, pp. 1–6. 2024. https://doi.org/10.23919/SoftCOM62040.2024.10721828

Rachel Van Campenhout, Ed.D., holds a bachelor of arts in philosophy and English from Duquesne University, a master of arts in digital publishing and writing from Emerson College, and a doctorate of education in instructional technology and leadership from Duquesne University. Starting as a learning engineer at Acrobatiq, Dr. Van Campenhout now leads the scholarly research and publication efforts of the VitalSource learning science team, covering topics that include the doer effect, automatic question generation, adaptive learning, and instructor implementation. Dr. Van Campenhout is also an active member of the IEEE IC Industry Consortium on Learning Engineering, was the co-chair of the 2023 Learning Engineering Conference, and has published papers on learning engineering and design, data science, and ethics in educational technology.

Benny G. Johnson, PhD, holds a bachelor of science in chemistry and mathematics from the University of Kentucky, and received his PhD in theoretical chemistry from Carnegie Mellon University, where he worked with a Nobel laureate. For the past fifteen years, Dr. Johnson has worked in the field of artificial intelligence for education, leading the research and development efforts of the Quantum technology for tutoring and assessment in chemistry, mathematics, accounting, and special education, and the machine learning predictive analytics technology of Acrobatiq, a Carnegie Mellon spin-off recently acquired by VitalSource Technologies. He leads VitalSource's research effort for using artificial intelligence to automatically create learning science-based courseware from textbooks at large scale. Dr. Johnson is the author of over fifty scholarly publications in academic journals and books and has delivered invited lectures at many national and international conferences. As principal investigator on various research and development projects, Dr. Johnson has received funding from the US Department of Education, National Science Foundation, National Institutes of Health, US Department of Energy, and Air Force Office of Scientific Research. He is a recipient of the Tibbetts Award, the highest recognition given by the federal government to small businesses for innovative research, and in 2007, he was inducted into the University of Kentucky's Alumni Hall of Fame.

Michelle Clark is the senior operations manager for the learning science team and supports new and ongoing research with our institutional partners. She has been with VitalSource for eight years in a variety of roles. She holds a BA in English and an MA in Digital Communication, both from University of North Carolina.

Dr. Melissa Deininger has been with Iowa State University since 2009. A specialist in the long century in France, her research includes publications on the French Revolution, Marie-Antoinette, the Revolution in pop culture, the Marquis de Sade, and the use of sound imagery and architecture in literature. She teaches courses in French culture and International Studies.

Shannon Harper is an assistant professor in the Department of Sociology and Criminal Justice, as well as core faculty in the U.S. Latino/a Studies Program at Iowa State University. She has a Ph.D. in Criminology, Law, and Justice from the University of Illinois Chicago and a Master of Public Administration with an emphasis on domestic violence from the University of Colorado Denver. Dr. Harper's research explores the relationship between intimate partner violence (IPV) and intimate partner homicide (IPH) and the gendered, structural, and cultural contexts through which both correlate and occur. Dr. Harper's work also investigates how race/ethnicity, class, gender, and other identities intersect to shape marginalized survivors' IPV and IPH experiences and interactions with the criminal legal system, specifically Latina and African American women. She has been published in multiple high-ranking, peer-reviewed journals, including the Journal of Interpersonal Violence, Feminist Criminology, Criminal Justice Studies, Homicide Studies, and Public Understanding of Science.

Kelly G. Odenweller is an Associate Teaching Professor in the Communication Studies program at Iowa State University. She received her B.A. in Communication from the University of Pittsburgh at Johnstown (2004) and her M.A. (2011) and Ph.D. (2015) in Communication Studies from West Virginia University. Kelly teaches courses focused on family, interpersonal, and professional communication, as well as Introduction to Communication Studies and the Senior Capstone Research Seminar. She researches how communication within and about families can socialize its members and foster social change for men and women. Her research has been published in *Journal of Family Communication*, *Communication Studies*, and *Southern Communication Journal.*

Erin Wilgenbusch, MA APR, is an accredited public relations professional and a teaching professor at the Greenlee School of Journalism and Communication. She has been teaching and serving as an academic adviser since the fall of 2002. She is also the faculty adviser to the Barbara Riedesel Iverson chapter of the Public Relations Student Society of America (PRSSA) at Iowa State University. Additionally, Erin serves on the Faculty Senate and several committees within the Greenlee School.