AI-Generated Practice for Textbooks: An Exploratory Analysis From the Classroom

Rachel Van Campenhout, VitalSource, United States
Michelle Clark, VitalSource, United States
Benny G. Johnson, VitalSource, United States

Abstract

Artificial intelligence has made it possible to generate high quality formative practice questions for use in higher education digital textbooks. Adding these automatically generated questions as a study feature for textbooks in an e-reader platform made it possible to democratize the learn-by-doing approach known to increase learning. Faculty in three different courses at a major public university in the United States assigned the automatically generated practice as a completion homework assignment with the textbook reading. In this paper, we investigate four automatically generated question types as well as an AG multi-question scaffolded tutorial using data from these three courses to better understand two research questions: how did these questions perform for students, and how did students choose to use them during their course? Additionally, survey data was collected to identify how students generally perceived the AG practice. Artificial intelligence can lead to unprecedented advances for teaching and learning technologies, but it is necessary to investigate how these tools perform for students in real-world contexts. The analyses from these classroom examples provide insights into how artificial intelligence can further benefit students in their everyday learning contexts.

Keywords: Automatic Question Generation, Textbooks, Student Perception

Introduction

Of the many uses for artificial intelligence in education, one high value application is automatic question generation (AQG). Formative practice is a well-known high-utility learning method that benefits students of all ages, and struggling students in particular (Black and William, 2010). Integrating formative practice within expository content in a learning by doing approach has been shown to be more effective for learning and increases learning outcomes (Koedinger et al., 2016; Carvalho et al, 2017). Research on AQG has increased in recent years with a wide application of uses (including the generation of formative practice) and varying methods for development, yet few research studies have evaluated AQG systems using student data Kurdi et al., 2020). Recent studies of automatically generated (AG) formative practice in courseware found that the AG questions performed as well as human-authored questions for engagement, difficulty, persistence, and discrimination when analyzing student data (Van Campenhout et al., 2021, Johnson et al., 2022).

This study extends the research on AI generated formative practice by investigating AG questions placed alongside digital textbook content for students to answer as they read. As examined in previous research, this study includes matching and fill-in-the-blank (FITB) questions, but also includes multiple choice (MC), submit and compare (S&C), free response, and the first known investigation of automatically generated scaffolded tutorial activities. This initial research focuses on three courses at two four-year public universities where the instructors incorporated the AI generated practice as part of the course assignments. The goals of this first investigation into the use of AG questions in a textbook learning resource are to 1) determine how the questions performed using student data, 2) explore how students used the AG practice in their learning context, and 3) explore student perceptions of the questions as a learning tool.

Methods

The AQG process for these questions uses the textbook as the corpus for natural language processing (NLP) and machine learning (ML) methods. Using Kurdi et al.'s (2020) AQG categorization, both syntactic and semantic approaches are used for the levels of understanding while the procedure for transformation is primarily rule-based. Essentially, the NLP and ML tools identify important sentences and key terms then transform them into comprehension questions while also sourcing multiple types of immediate feedback. This AQG system also includes multi-question tutorial activities. The AG questions are delivered next to the digital textbook interface as a pop-up panel so students can answer questions and read the text simultaneously.
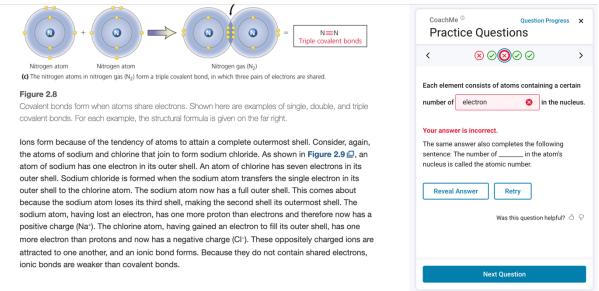
Figure 1: An example of an e-textbook page with the questions open in a side panel. The FITB question is shown with scaffolding feedback, and options for further interactions such as reveal answer, retry, and providing a rating.

In the Fall of 2022, three instructors volunteered to incorporate the AG practice into their course where the textbook was assigned as the primary learning resource. For Cognitive Psychology, the instructor assigned 4% of the student's grade for completion of a minimum of 80% of the practice, due by the end of the course. Psychology had 115 students, was online with synchronous sessions, and had exams as the primary graded component, with discussions, homework, and attendance included. The Sociology course used a Women, Gender, and Crime textbook and the instructor assigned 10% of the course grade to the AG practice, expecting students to complete the chapter reading and practice before coming to class in-person weekly. This class had 50 students and written assignments and projects constituted the majority of the graded components. The final course was a Public Relations (PR) capstone course for seniors, and the instructor assigned about 15% of the students' grade for completing 80% of the practice. The 31 students were expected to read the text and do the practice weekly for the first 8 weeks of the course, with the remainder of the time dedicated to a final project. Instructors had weekly data reports to monitor student engagement and apply points and also sent students a short, anonymous survey after students completed their work.

Results

Investigating performance metrics is a quantitative method of evaluating automatically generated questions that is easily comparable across research studies (Kurdi et al., 2020). In each course, students were assigned a minimum of 80% of practice completion to earn points in the course. Because of this incentivization, engagement with the practice was high across all courses. In Sociology students completed a mean of 86% of assigned practice with 50% of students completing 100% of practice in assigned chapters. In Psychology, students completed a mean of 92% of assigned practice with 68% of students completing all assigned practice. In PR, students completed a mean of 90% of assigned practice and 38% of students completed 100% of all assigned practice.

Course	Metric	Matching	FITB	MC	S&C
Sociology	N	32	109	4	10
Sociology	Mean Difficulty	0.92	0.81	0.88	0.93
Sociology	Mean Persistence	1.00	0.98		
Psychology	N	36	140	8	9
Psychology	Mean Difficulty	0.66	0.63	0.49	0.94
Psychology	Mean Persistence	0.95	0.92		
PR	N	59	141	10	14
PR	Mean Difficulty	0.79	0.86	0.57	0.89
PR	Mean Persistence	0.87	0.72		

Table 1. Mean difficulty and persistence for each question type for each course.

Table 1 reports the mean score on the first attempt and mean persistence (percentage of students initially answering incorrectly who continued until answering correctly) for the question types. For both calculations, only questions with 25 or more responses were included, and persistence was only included for matching and FITB, as MC and S&C had fewer questions and also often fewer students initially answering incorrectly. Sociology had the highest mean scores and also the highest persistence while Psychology had the lowest mean scores yet maintained persistence over 90% for both question types. PR had mean scores in the middle of the range but lower persistence means. It is notable that the S&C questions were open response questions, so mean scores in these cases are calculated by students voluntarily rating their response as correct or incorrect after submitting and reading the expert answer. Free response questions were not included in this table as they are not scored.

In addition to the AG questions that exist as stand-alone items in an activity, tutorial-style questions were also automatically generated. These tutorials are based on specific generated questions and automatically trigger a set of questions or interactions for students based on their response. While tutorials are less common than regular activities, they constituted 10.3% of all data in Sociology, 10.5% of data in Psychology, and 12.1% of data in PR. One tutorial type is attached to MC questions. Once students answered, they were told that another student selected one of the incorrect responses and asked to help that student and type a response (a free response question). While the incorrect responses are anonymous and not from their peers in the course, this tutorial type gives students a metacognitive activity with a social element. For example, a multiple choice question in the Sociology course provides a definition and asks students to select the correct term. The correct response was Lifestyle Theory, and students who selected this option were then prompted with: "Another student answered Life Course Theory. What would you say to help them understand their error?" In total, 27 students responded to this AG tutorial question. Two responses were non-answers and two simply told the student the correct answer (e.g., "it's lifestyle theory you goofy goober"). The remaining 23 students entered explanations to help explain or differentiate Lifestyle Theory and Life Course Theory. One student explained, "Life style theory explains how an individual's life choices affect their victimization. The life course theory explains how all an individual's life events contribute to their victimization." While the initial multiple choice question provided feedback and a chance to retry for students who got the question incorrect, those who got the question correct were offered an additional way to apply their knowledge.

For students answering the FITB questions, there is an additional possible interaction—spelling validation. A spelling mistake or keyboard error would normally render responses incorrect, which would mean the student is not actually evaluated on their knowledge of the content. Using an edit distance calculation, any misspelled response close to the correct answer triggers a suggestion box with possible corrections, including the correct term, allowing the student to select the term they intended before. The answer suggestion rate per course is relatively low (Sociology = 3.4%, Psychology = 2.9%, PR = 3.1%). However, this feature becomes more meaningful when looking at students who received help. A student in Sociology received 30 answer suggestions, meaning, they provided 30 misspelled responses close to the correct response (for example, "canidate," "critisicm," and "philanthrophy"). Instead of being marked incorrect for these responses, the spelling validation tool gave this student the opportunity to identify and submit the term they intended to use. This feature can be very meaningful to students who know the content but may be poor spellers or typists.

Student surveys were distributed by the course instructors. The instructors for Psychology and PR gave students a few minutes in class to complete the survey while the Sociology instructor distributed asynchronously via email. All student responses were anonymous and voluntary. Psychology was the largest class with 115 students and 55% of students responding. PR was the smallest class with 31 students, yet 87% responded. Sociology had 55 students and a 28% response rate, which may be lower due to the difference in survey delivery method. While some questions asked are not reported here for brevity, Table 2 compares student responses from all three courses.

Question	Course	N					
Generally, how do you like using digital textbooks as a learning resource compared to print books?			Much Better	Somewhat Better	About the Same	Somewhat Less	Much Less
	Psych	63	31.7%	30.2%	15.9%	12.7%	9.5%
	PR	27	33.3%	29.6%	11.1%	25.9%	0%
	Soc	44	42.9%	28.6%	7.1%	14.3%	7.1%
How important do you think reading the textbook was to your overall success?			Extremely	Very	Moderately	Somewhat	Not at all
	Psych	63	6.3%	38.1%	46.0%	7.9%	1.6%
	PR	27	14.8%	37.0%	22.2%	25.9%	0%
	Soc	14	0%	28.6%	57.1%	14.3%	0%
In general, do you think doing practice questions while reading is helpful for learning?			Yes	Maybe	No		
	Psych	63	68.3%	22.2%	9.5%		
	PR	27	81.5%	14.8%	3.7%		
	Soc	14	64.3%	28.6%	7.1%		
How helpful did you find the practice questions for studying?			Very helpful	Moderately helpful	Somewhat helpful	Not at all helpful	
	Psych	53	32.1%	32.1%	32.1%	3.8%	
	PR	27	37.0%	40.7%	18.5%	3.7%	
	Soc	14	14.3%	50.0%	35.7%	0%	
How helpful were the practice questions in preparing you for the course assignments and exams?			Very helpful	Moderately helpful	Somewhat helpful	Not at all helpful	
	Psych	53	24.5%	35.8%	32.1%	7.5%	
	PR	27	29.6%	37.0%	29.6%	3.7%	
	Soc	14	7.1%	50.0%	42.9%	0%	

Table 2. Survey questions and responses for all courses.

Some questions were included to get a baseline of students' feelings on digital learning resources in general before investigating the practice questions. Results showed that across

courses there was a percentage of students (21.4–25.9%) who liked digital textbooks less than print books. Students self-reported their reading frequency which generally matched the reading frequency in the course data, but not all students felt the textbook was equally necessary for success in the course. While some students felt the textbook was extremely or very important, some felt it was only somewhat important.

When asked if they felt doing practice while reading was generally beneficial for learning, most students said yes (64.3–81.5%), some said maybe (14.8–28.6%), and a few said no (3.7–9.5%). Psychology was the only course where any students responded that they seldom or never did the practice and those 10 students were asked a different set of follow-up questions. When considering how helpful students found the practice questions for studying, 64.2–77.7% thought they were moderately to very helpful. No students in Sociology thought they were unhelpful, while 3.8% and 3.7% of students in Psychology and PR found them unhelpful. For helpfulness for preparing for assignments and exams the percentages only minimally changed.

Conclusion

This initial investigation gives the first glimpse into how automatically generated questions were used as a study tool with the digital textbook by students in a classroom setting. While instructors expect students to read the textbook, the addition of formative practice as a feature of the reading platform allowed them to incorporate that practice into their teaching and monitor student engagement. Adding a small percentage of points for completing the questions motivated most students to do them while they read, and many students completed 100% of the available practice—more than the 80% required. The data from these questions revealed differences in question difficulty between Sociology and Psychology, yet both had very high persistence. PR presented an interesting case where the difficulty was mid-range, yet it had the lowest persistence which provides cause for further investigation. Other data also revealed the support that spelling suggestions gave to select students. Another encouraging finding is the high engagement with the automatically generated tutorial series of questions. Generating question types for students to engage in metacognitive reasoning expands their learning opportunities beyond typical recognition or recall question types.

Student perception is important to gauge, especially for learning tools developed using artificial intelligence. Survey responses revealed that there is a small percentage of students who prefer print books over digital, do not believe doing practice is helpful for learning, and did not find the automatically generated practice helpful. It is reasonable to expect some students may feel this way. However, the majority of students preferred digital textbooks, thought doing practice while reading was beneficial for learning, and found the practice questions helpful for studying and preparing for assignments and assessments. Student perception of the automatically generated questions were generally positive. While the surveys were anonymous, respondents were self-selected. One limitation was the response rate, particularly for Psychology and Sociology which could impact the results. By contrast, PR had nearly all students respond and some of the most positive perceptions of the practice. It is also noteworthy that the Psychology student survey was delivered prior a large volume of engagement before the last exam. In future research, it would be beneficial to maximize student responses for all courses and deliver the survey after students have finished studying for the term.

This exploration into how students engaged with AG practice is promising and leads to several possible avenues of continued research. An immediate next step would be to investigate the performance of the AI-generated questions on a larger scale to analyze all question types on performance metrics. Secondly, a more in-depth investigation of the tutorials and how they support student learning would be another meaningful extension for AQG research in this area. Lastly, research on how formative practice alongside textbook content benefits student learning, both qualitatively and quantitatively, should be reported across different teaching and learning contexts.

Acknowledgements

References

Black, P., & William, D. (2010). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan 92*(1), 81–90. https://doi.org/10.1177/003172171009200119

Carvalho, P. F., McLaughlin, E. A., & Koedinger, K. R. (2017). 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.), 204–209.

Johnson, B. G., Dittel, J. S., Van Campenhout, R., & Jerome, B. (2022). Discrimination of automatically generated questions used as formative practice. *Proceedings of the Ninth ACM Conference on Learning@Scale, 325-329.* https://doi.org/10.1145/3491140.3528323

Koedinger, K., McLaughlin, E., Jia, J., & Bier, N. (2016). 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. http://dx.doi.org/10.1145/2883851.2883957

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education, 30*, 121–204. https://doi.org/10.1007/s40593-019-00186-y

Van Campenhout, R., Dittel, J. S., Jerome, B., & Johnson, B. G. (2021). Transforming textbooks into learning by doing environments: An evaluation of textbook-based automatic question generation. *Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education.* http://ceur-ws.org/Vol-2895/paper06.pdf