# Advancing Intelligent Textbooks with Automatically Generated Practice: A Large-Scale Analysis of Student Data

Rachel Van Campenhout, Michelle Clark, Bill Jerome, Jeffrey S. Dittel, and Benny G. Johnson

*VitalSource Technologies, Raleigh, NC, USA*

**Abstract**

Integrating formative practice questions with textbook content at frequent intervals creates an active learning environment that is more effective for student learning. Advances in artificial intelligence have made it possible to develop automatic question generation systems robust enough for use with students at scale. In this paper, we analyze five types of automatically generated questions using data from hundreds of thousands of students across more than eight thousand textbooks. The difficulty and persistence performance metrics of these questions build on previous research and reveal insights into question performance and student behavior. Metacognitive tutorial activities are also generated, and investigation into students' open-ended responses show differences in how students apply what they have learned from the text.

**Keywords**

Artificial intelligence, automatic question generation, formative practice, textbooks, learning by doing

## 1. Introduction

The use of artificial intelligence for automatic question generation (AQG) has become an increasingly viable option for incorporating learning by doing in digital textbooks. Textbooks are passive learning resources, and reading has been shown to be less effective for learning than the combination of reading and practice. Rereading is regarded as a low utility study strategy compared to alternative active learning approaches [1]. Incorporating formative practice questions with text content at frequent intervals in a learning by doing approach has been shown to have approximately six times the impact on learning than merely reading [2][3]. This active learning method fosters the doer effect, a learning science principle proven to be causal to learning [3][4][5]. Furthermore, the doer effect remains significant even after accounting for student prior knowledge and individual characteristics [2][6], confirming this learning by doing approach is beneficial for all students. Consequently, the implementation of automatically generated formative practice in digital textbooks provides an efficient means of drastically scaling the doer effect for students.

Recent years have seen a surge in research on AQG systems across a variety of educational applications. However, few studies have evaluated automatically generated (AG) questions using student data [7]. In our previous work, an AQG system was developed that uses electronic textbooks as the corpus for natural language processing and machine learning. Initially, two question types (fill-in-the-blank and matching) were generated and used in automatically generated courseware [8]. These AG questions were placed alongside human-authored questions, and research on students' interactions with these questions in natural learning contexts focused on several key performance metrics: engagement, difficulty, persistence, and discrimination [9][10]. Findings indicated no meaningful differences between AG and human-authored questions. Instead, differences were observed in relation to the cognitive process dimension of the question type. For example, fill-in-the-blank questions correspond to a recall cognitive process, whereas matching questions involve a recognition cognitive process [11].

*iTextbooks'23: Fifth Workshop on Intelligent Textbooks,* July 03, 2023, Tokyo, Japan
EMAIL: rachel.vancampenhout@vitalsource.com (A. 1); benny.johnson@vitalsource.com (A. 5)
ORCID: 0000-0001-8404-6513 (A. 1); 0000-0003-4267-9608 (A. 5)

The AG and human-authored questions were similar to each other based on their cognitive process dimension.

In this work, we expand on this existing research by investigating the performance of AG questions that were incorporated into digital textbooks as a free added learning feature, named CoachMe, within VitalSource's Bookshelf e-reader platform. While this is a different learning environment than courseware, adding formative practice directly to the textbook reading experience allows for immediate scaling of learning by doing. Since 2022, more than 2.3 million AG questions have been placed in over 8,000 textbooks, making them available to any learner using those textbooks in any context. In addition to a larger scale than previous research [9], this study includes AG question types not previously investigated, such as multiple choice, free response, and submit-and-compare. Furthermore, certain question types were combined to create multi-step "tutorial" activities, in which a student's response to a question can trigger additional follow-up questions using simple branching conditional logic. Of particular interest are AG metacognitive free response questions employed in some tutorial types (described below), as they represent the first questions generated by this AGQ system intended to evoke the "understand" and "evaluate" cognitive process dimensions [11] with a metacognitive component. The goal of these tutorials is to take advantage of the learning benefits from self-explanation and metacognitive thinking [12]. While the majority of the AG questions are still recognition and recall types and on the lower levels of Bloom's Taxonomy, the metacognitive tutorials offer an opportunity for students to engage in higher Bloom's levels.

This study evaluates the AG questions in two different ways: analyzing questions based on difficulty and persistence where applicable, and performing an initial analysis of students' textual responses. There are two primary research goals in this investigation:

1. To learn more about the performance of AG questions at a large scale.
2. To gain a better understanding of emerging patterns in student behavior.

Additionally, a comparison is made between the data set of all student-question interactions to date and a single university course using the AG questions wherein the student learning context is known. As artificial intelligence is increasingly used for educational applications such as AQG, it is increasingly necessary to report and establish benchmarks for basic performance metrics using student data. Through these analyses, this study aims to provide valuable new insights into how AQG can help transform passive textbooks into interactive learning by doing environments to support student learning.

## 2. Methods

Following the recommendation of Kurdi et al. [7], we report the essential information summarizing the AQG method used. The purpose of AQG is to generate questions for formative practice as students read a textbook. Generation uses both syntactic and semantic levels of understanding, and an expert-developed rule-based approach is used for the procedure of transformation. The textbook is the input corpus for the natural language processing methods. When possible, feedback for incorrect answers is also generated from the textbook content. The AQG method is not designed for a specific domain and is applicable to a wide variety of subject matter; however, the method is not applicable for some domains, mainly mathematics and language learning. As seen in Figure 1, the questions open in a panel next to the textbook content, allowing students to refer back to the content if needed while they answer.

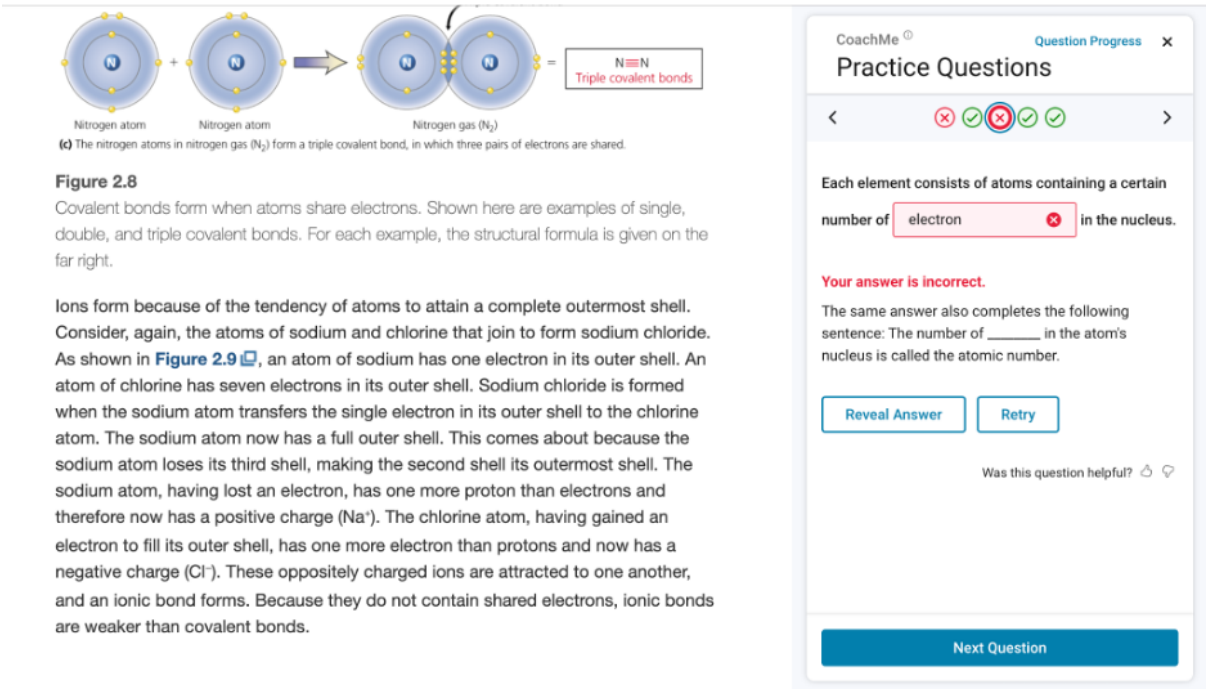


Figure 1. An example AG practice question in a chemistry textbook.

The question types in this study include cloze questions created from important sentences in the textbook content: fill-in-the-blank (FITB) questions with a single answer blank (recall), and matching questions with three answer blanks (recognition). Important sentence identification is done by ranking the sentences in each textbook section with the TextRank algorithm [13]. Several considerations factor into selection of answer word(s) within sentences, including part of speech (must be noun or adjective) and frequency in the textbook corpus. Feedback is generated where possible using textbook sentences related to the question stem, such as a different sentence containing the same answer word (illustrated in Figure 1); outcome feedback is always available. These question types have been previously studied using student data from natural learning contexts [9].

New AG question types not studied in our previous work include multiple choice questions, in which a term or definition from the textbook's glossary is used as the question stem and the choices consist of the corresponding definition or term plus two to three distractors created from coordinate glossary terms; for example, a question with the correct answer "covalent bonding" might have "ionic bonding" and "hydrogen bonding" selected as distractors. Another new question type is self-graded submit-and-compare, in which the student is asked to write a short free response answer (approximately one sentence) to a question prompt. The student's answer is not automatically scored for this question type. Instead, after submitting the answer the student is given a model correct answer obtained from the textbook content for comparison, and then asked to self-report if their answer was correct. In this work, these questions asked the student to write a definition for a given textbook glossary term, and the model answer was the textbook's definition.

The new multi-step tutorial activity type has been utilized for various purposes, including scaffolding, remediation, and enrichment. Examples include following an incorrect response to a question with an easier version of that question instead of simply having the student try the same question again or revealing the correct answer, creating flashcards from content of missed questions, and following a correctly answered question with a more advanced question on the same topic. In Section 3.2, a particular type of enrichment tutorial, in which the student is asked to explain an incorrect answer to a question they have just answered correctly, is the focus of analysis.

In order to achieve our research goals, this paper focuses on an exploratory data analysis of a historical data set. A benefit of digital learning environments is the large volume of microlevel data collected [14]. In this case, each interaction students had with the automatically generated questions was recorded, including each answer attempt and correctness state. This large data set of interaction

events can reveal new insights into the performance of the AG questions by exploring descriptive statistics for each question type.

## 3. Results

The student-question interactions for this analysis were aggregated across all textbooks with CoachMe questions in use from January 4, 2022 (launch) through May 12, 2023, when most courses using CoachMe for the Spring 2023 term had completed. This includes 8,407 textbooks, 334,902 students, 941,318 unique questions, and 8,753,453 interaction events. While the number of unique questions answered is just short of a million, the total number of questions answered is 5,370,981. These differ because multiple students can answer the same question. Furthermore, since these questions are used as formative practice, students can answer each one many times (for example, until they get it correct). The total number of individual answer attempts is 7,077,271. This is the largest data set of students answering AG questions analyzed in AQG research known to date.

### 3.1. Difficulty and Persistence

The first validation check for the AG questions is the difficulty and persistence performance metrics. Question difficulty is important to monitor, as research has shown that questions that are too easy or difficult could deter students [15]. It is also important to note, though, that common difficulty index boundaries set in the literature are regarding assessment items for high-stakes exams, not formative practice. Unlike in a summative assessment context, a greater variety of factors may influence observed difficulty for low-stakes formative practice (for example, students have the option of searching for answers if desired). There is no set standard for difficulty for formative items used for learning by doing. However, prior research on AG and human-authored questions used as practice in courseware environments found very difficult questions had much lower persistence (students stopped trying before reaching the correct answer) while the easiest questions maintained high persistence rates [9].

#### 3.1.1. Difficulty

In this study, difficulty is determined by the students' first answer attempt on the questions. The difficulty index (percentage of students answering correctly on the first attempt) for each AG question type is noted in Table 1; note that the definition of difficulty index means higher values correspond to less difficult questions. Using a two-tailed $z$ test, all differences in difficulty index by question type are statistically significant ($p << 0.001$). The difficulty of the recognition question types, matching (79.3%) and multiple choice (72.3%), aligns with previous research findings [9]. Compared with results from six courseware environments, the AG matching was one of the easier question types with difficulty index ranging from 81% to 90% [9, Table 5]. Given that the courseware studied was used in specific classroom contexts, the similar trend for AG matching across a larger aggregated data set is a positive finding. In the previous research, human-authored multiple choice questions were included in the analysis, which ranged from the mid-sixty to mid-seventy percent difficulty. While the multiple choice questions here are automatically generated, the mean of 72.3% falls within the range of the human-authored questions. In this case, it is also consistent that the matching questions are easier than the multiple choice.

The FITB are, by contrast, a recall question type and are the most difficult at 54.7%. This finding is consistent with previous research, wherein the difficulty for AG FITB were largely in the 60% range [9, Table 5]. A lower first attempt accuracy in this case seems reasonable given the additional effort required for the recall type (to determine and correctly input a response) and the aggregation of a more context-varied data set.

For the self-graded submit-and-compare question type, the difficulty index of 80.5% was obtained from student self-ratings. This indicates that students likely responded honestly in reflection on their own answers. The last AG question type was free response, which was not scored so a difficulty index could not be evaluated.

**Table 1**
Difficulty index by AG question type.

| Question Type | Total Answered | Difficulty Index |
|---|---|---|
| Matching | 1,630,654 | 79.3 |
| Multiple Choice | 85,881 | 72.3 |
| FITB | 3,514,483 | 54.7 |
| Self-Graded Submit-and-Compare | 108,691 | 80.5 |
| Free Response | 31,215 | NA |

### 3.1.2. Persistence

As these questions are formative, if a student responds incorrectly at first, they can continue to answer. Persistence is the rate at which students continue to answer until they reach the correct response. While the persistence data set is a subset of the difficulty data set—as it includes only students who answer questions incorrectly on their first attempt—persistence is a separate metric focused on the student's decision to persevere on an incorrectly answered question. While persistence is not entirely independent from the difficulty of the question, as that may influence a student's decision to persist, it is also likely that persistence is influenced by the cognitive process type of the question (as seen in [9]) or the student's own motivation. Persistence is a valuable metric to monitor for two reasons. First, VanLehn [16] describes the importance of persistence on the learning process itself. Second, persistence by question type can help indicate the performance of the AG questions as a learning tool.

Table 2 gives the persistence rate for each question type. The highest persistence rate was observed for multiple choice (93.6%), which also happens to require the least effort to answer. The next highest persistence rate was for matching (69.5%), followed by FITB (58.5%). The differences in persistence by question type are statistically significant ($p << 0.001$). In previous research, persistence rates exceeded 80% for AG FITB and matching question types [9]. However, it is important to note that the course context could potentially have a larger impact on persistence than the varied contexts aggregated in this study, in the majority of which student engagement with the questions was optional.

It is also notable that the persistence rates are ordered by the cognitive process dimension [11]. The highest persistence rate was observed for multiple choice questions, a recognition type that also required the least effort to answer. Matching questions—another recognition type— resulted in the next highest persistence rate. The lowest persistence rate was on the fill-in-the-blank (FITB) questions, which are a recall type.

**Table 2**
Persistence rate by AG question type.

| Question Type | Incorrect First Attempts | Persistence |
|---|---|---|
| Matching | 335,535 | 69.5 |
| Multiple Choice | 23,620 | 93.6 |
| FITB | 1,580,743 | 58.5 |

### 3.1.3. FITB "Non-Genuine" Answers

The FITB questions provide an opportunity to investigate student behavior more closely, as the answers students input are recorded. These responses provide a large data set of highly varied qualitative data. Due to the design of the user interface, students must attempt the question before they receive any other information. It is possible for students to input a non-genuine answer attempt, i.e., one that is not convincing as an attempt to enter a word. In order to explore how often this might be occurring, FITB responses were analyzed with a set of simple rules developed to estimate the percentage of non-genuine first attempts. These rules included categories such as:

- Very short answers (less than 3 characters)
- Answers with no vowels
- Answers containing punctuation
- Known common non-answers (e.g., “idk”)

While not every non-genuine answer may be identified using these rules, it provides a good estimate for how often students applied this approach.

Of the nearly 1.6 million incorrect first attempts for FITB questions, 12.2% were categorized as non-genuine answers. Of these, the majority (82.7%) were very short answers. The most common short answer was “d”, which was entered 9,993 times. The commonly used response “idk” was entered 5,408 times. To investigate further, for this group of non-genuine first attempt responses a persistence rate of 46.5% was identified. Of the students who employed this strategy to answer the FITB questions initially, nearly half followed this by ultimately completing the question correctly.

While we cannot know all the reasons for this behavior, two possibilities warrant discussion. The first is that some students may not be taking the practice seriously and therefore apply minimal effort. Rather than come up with an answer, they type in a non-genuine answer according to a selected strategy. That approach would certainly not shock most educators, who are well aware that not all students share the same motivations or strategies for learning. However, a second possibility for this behavior must also be considered. Because the user interface requires an attempt before providing feedback or the option to reveal the correct answer, if a student genuinely does not know how to answer, they might submit a non-genuine response to get to the feedback. In this case non-genuine answers may be less driven by a student’s chosen strategy, but rather as a response to a difficult question. Given this finding, potential improvements to the user interface is a topic for future investigation.

### 3.1.4. Single-Course Comparison

The results thus far encompass all student-question interactions with the CoachMe AG questions across all learning contexts. By comparison, there were several university courses that assigned these questions as homework, therefore driving student engagement with all questions. One such course was a criminal justice course conducted at a major public university in Fall 2022 with 50 students. The instructor assigned the AG questions in the course’s textbook [17] for homework on a weekly basis. Students received ten points per chapter for completing a minimum of 80% of the practice questions. Points were awarded for completion only, not accuracy.

Table 3 shows total questions answered, difficulty index, and persistence rate for each question type. The results for this particular course are in contrast to the aggregated results in Tables 1 and 2. In this course, all question types had a higher difficulty index by at least ten percentage points. The FITB questions in this course have a difficulty index of 78.0% compared to the aggregated 54.6%. Similarly, persistence in this course was over 90% for all question types. The results for AG matching and FITB are more similar to the previously examined results from a courseware environment [9, Table 5]. The rate of non-genuine answers was also calculated. At 17.1%, this course had a higher rate than the aggregated data; however, the persistence for non-genuine answers was 92.0%—essentially double the aggregated persistence rate.

**Table 3**
Mean difficulty and persistence by AG question type for criminal justice.

| Question Type | Total Answered | Difficulty Index | Persistence |
|---|---|---|---|
| Matching | 1,465 | 91.4 | 96.0 |
| Multiple Choice | 193 | 85.0 | 96.6 |
| FITB | 5,121 | 78.0 | 91.7 |
| Self-Graded Submit-and-Compare | 458 | 93.0 | NA |
| Free Response | 128 | NA | NA |

## 3.2. Metacognitive Tutorials

The tutorials provide a distinctive perspective on student learning experiences with the textbook content. Results from a representative tutorial type are analyzed here: a metacognitive activity triggered by answering an initial multiple choice question correctly. When providing the correct answer on the first attempt, students were told that another student had selected one of the incorrect distractor responses and asked to help that student by giving an explanation of the error as a free response. An example is shown in Figure 2. While the incorrect response did not originate from an actual peer in the course, this tutorial type gives students a metacognitive activity with a perceived social aspect.

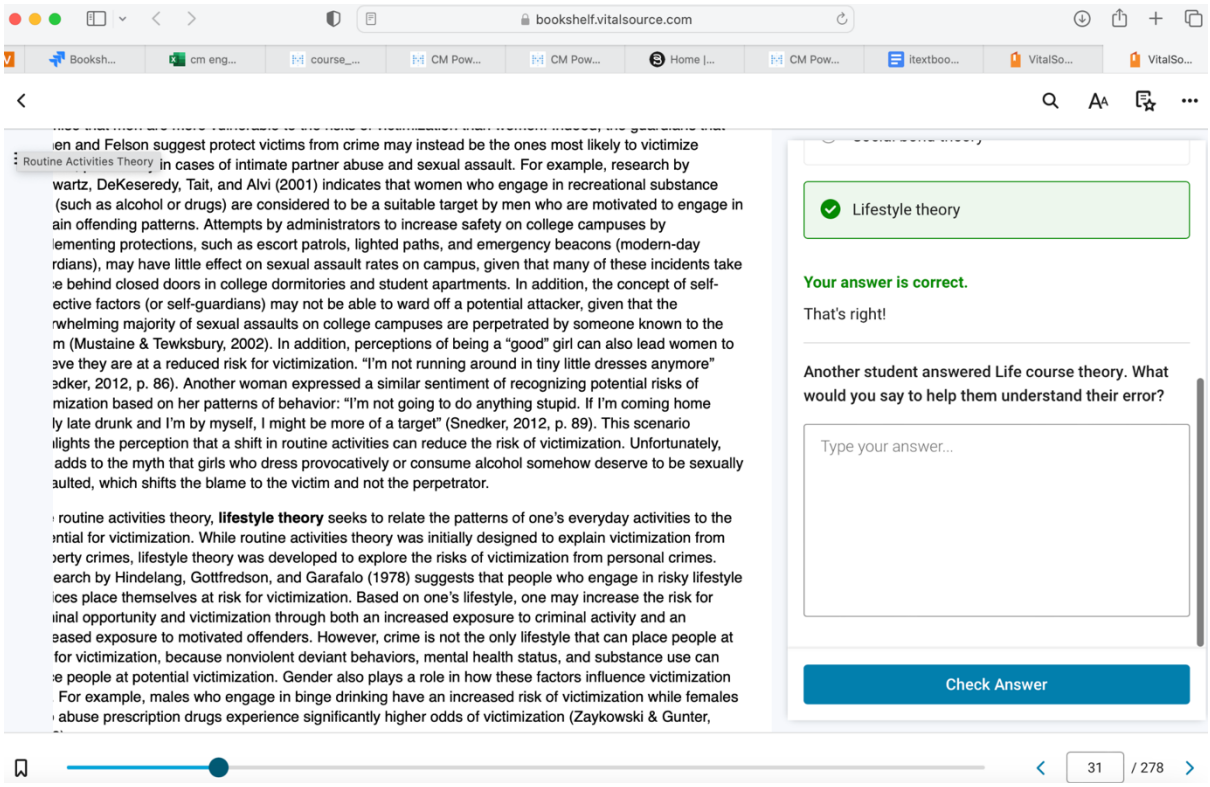


Figure 2. A metacognitive tutorial follow-up question from criminal justice.

As a new type of automatically generated activity designed to elicit more complex reasoning and deeper understanding of the textbook material, the analysis of this metacognitive tutorial type reveals new insight into the learning possibilities of AQG. In total, 30,577 answers to the metacognitive tutorial step were submitted.

### 3.2.1. Answer Length

The first approach to understanding how students responded was by examining answer length, shown in Table 4. The mean answer length was 12.5 words, which could potentially constitute a succinct yet complete sentence. The minimum response (as required by the user interface) is one word, while the 25th percentile is three words, which may not be a genuine attempt to address the prompt. Some examples of three-word answers are “i don’t know” and “read the book.” The 75th and 90th percentiles indicate longer sentences or multiple-sentence answers. The maximum lengths encompass tactics such as copying a section of the textbook into the answer text box. To estimate the percentage of non-genuine answers, answers with less than five or greater than 200 words were used, resulting in 30.7% of answers deemed non-genuine.

**Table 4**
Descriptive statistics for metacognitive tutorial answer length in words.

| | |
|---|---|
| Mean | 12.5 |
| Standard deviation | 12.3 |
| Minimum | 1 |
| 25% percentile | 3 |
| 50% percentile | 10 |
| 75% percentile | 18 |
| 90% percentile | 28 |
| Maximum | 421 |

The distribution of answer lengths, shown in Figure 3, has an interesting shape. The most prominent peak is observed for responses under five words—the criterion established for identifying non-genuine short answers. Yet there is a plateau evident between eight and ten words, resulting in an almost bimodal shape. The graphical representation shows the presence of multiple distinct behaviors pertaining to the length of students' answers.

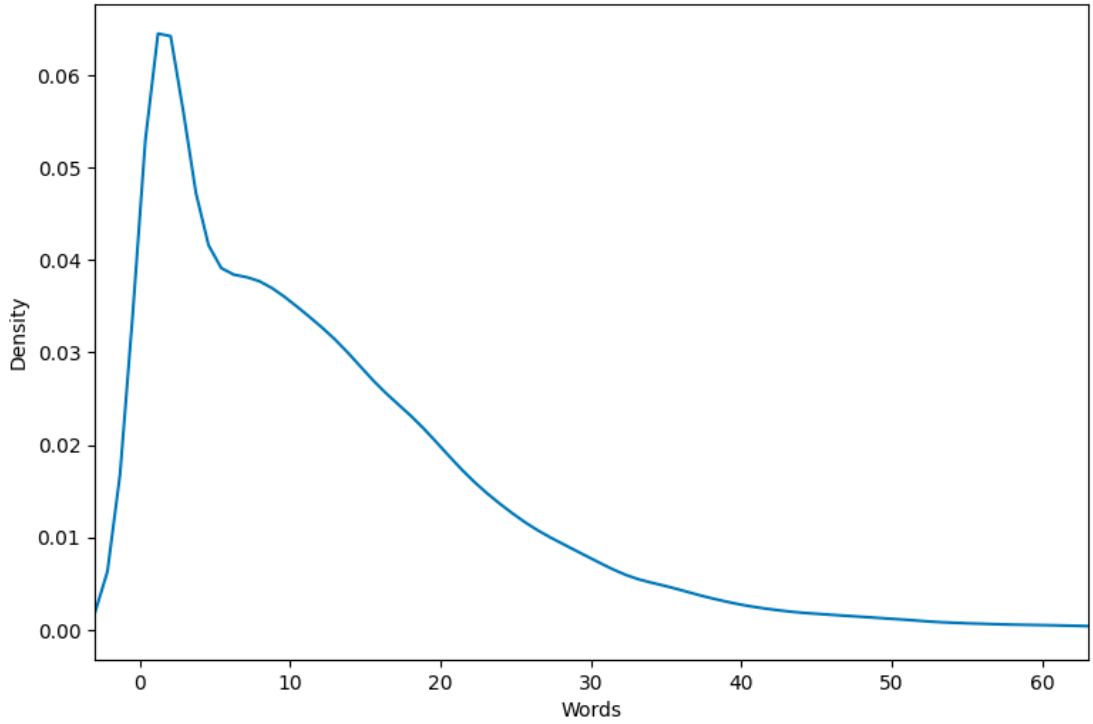


Figure 3: Distribution of answer lengths for metacognitive tutorial free response questions.

## 3.2.2. Key Term Analysis

To analyze the student responses from a different perspective, the correct answer and distractor key terms from the multiple-choice-to-metacognitive-prompt tutorial activity were utilized. In this scenario, the students must first select the correctly defined term, which is then followed by an incorrect distractor term to explain to a peer. This specific setup allows for categorizing student responses based on the usage of these key terms in their answers. There are four possible categories: neither term, distractor term, correct term, both terms. For the 9,158 tutorials of this type that were completed, the distribution of key terms usage is presented in Table 5. The largest category is neither term included, which comprises 43.6% of responses—an even larger percentage than for non-genuine short answers. The next largest groups are students who included only one of the key terms, either the distractor (23.5%) or correct (18.4%) term. The smallest group was students who included both terms (14.6%).

**Table 5**
Percentage of answers containing distractor/correct key term combinations.

| | | Distractor Term | |
|---|---|---|---|
| | | No | Yes |
| Correct Term | No | 43.6 | 23.5 |
| | Yes | 18.4 | 14.6 |

As there are distinct categories based on the key terms included in the answers, the distribution of answer length was plotted for each category, shown in Figure 4. This demonstrated notable differences among the categories. A Kruskal-Wallis H test showed that the answer length mean ranks in the categories are not all the same ($p << 0.001$). The group that did not use either term had a peak answer length of approximately three words, followed by a steep decline by ten words. Although some students gave longer answers without using either term, it is seen this group contains the majority of the non-genuine answer attempts. For the groups that used only one term (either the distractor or the correct term), a peak answer length was observed near ten words, which also corresponds to the 50th percentile of all answer lengths from Table 4. These groups exhibited a more gradual decline in answer length frequency. In contrast, the group that used both terms had a peak length of over 20 words and consistently longer answers than the other groups. While other useful ways to evaluate the responses may exist, classification by the distractor and correct key terms from the metacognitive tutorial prompts reveals distinct groupings that could potentially be used in the future for delivering more tailored feedback, for example.

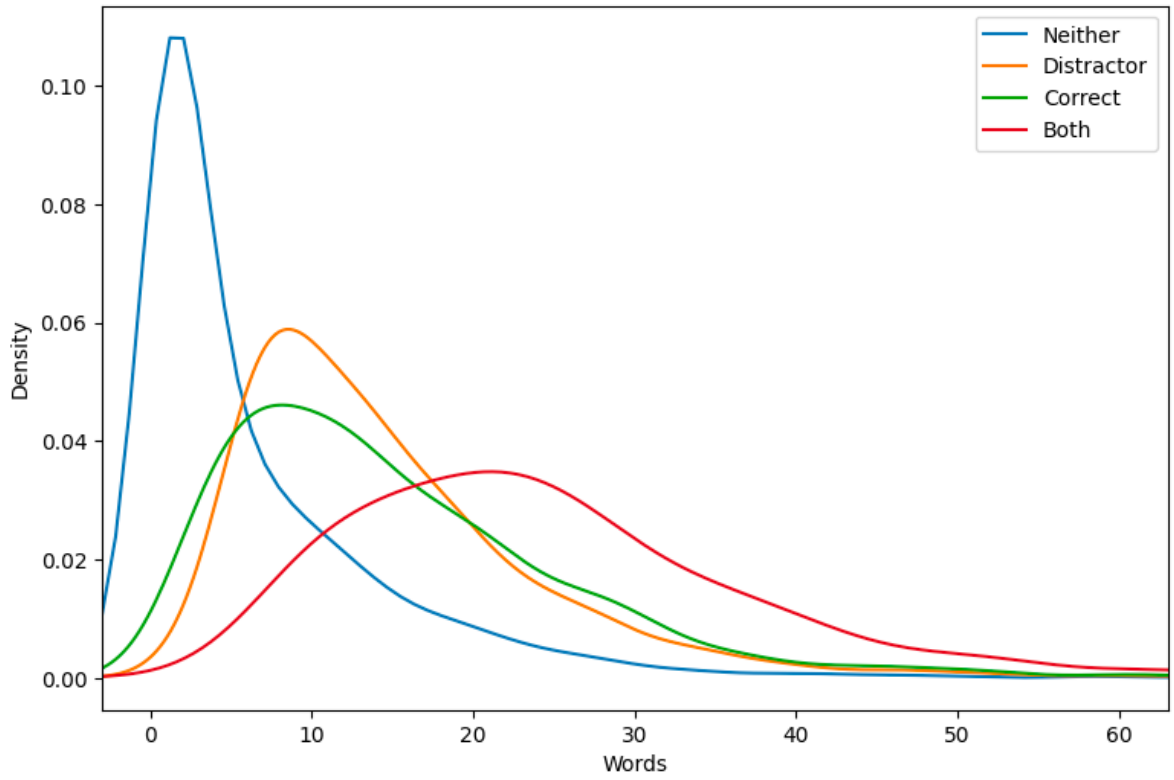


Figure 4. Answer length distributions for each key term usage category.

## 3.3. Single-Course Comparison

In an example tutorial from criminal justice, a multiple choice question gave students a definition ("Developed to explore the risks of victimization from personal crimes and seeks to relate the patterns of one's everyday activities to the potential for victimization") where the correct response was "Lifestyle theory." While the initial multiple choice question provided feedback and a chance for the student to try again if they answered incorrectly, those who answered correctly were offered an additional opportunity to apply their knowledge. These students were then prompted with: "Another student answered 'Life course theory' [a distractor from the initial question]. What would you say to help them understand their error?" A total of 27 students responded to this AG tutorial question. Two responses were non-genuine answers, and two others simply provided the correct answer (e.g., "it's lifestyle theory you goofy goober"). The remaining 23 students submitted responses to help explain or distinguish between lifestyle theory and life course theory. For example, one student explained, "Life style [sic] 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." This response demonstrates the student's understanding of both the correct and incorrect terms by their attempt to differentiate between them.

The classification of answers by key term(s) referenced was analogously examined within the specific context of the criminal justice course. As shown in Table 6, 17.9% of answers did not contain either of the key terms, whereas 41.8% of answers incorporated both. Notably, these proportions contrast with those in the aggregated results from Table 5.

**Table 6**
Percentage of answers containing distractor/correct key term combinations for criminal justice.

| | | Distractor Term | |
|---|---|---|---|
| | | No | Yes |
| Correct Term | No | 17.9 | 16.4 |
| | Yes | 23.9 | 41.8 |

## 4. Conclusion

Recent advancements in artificial intelligence, specifically in natural language processing and machine learning tools, have facilitated the development of automatic question generation systems capable of producing high-quality formative practice questions. AQG systems can accomplish what is otherwise too costly—the generation of millions of formative practice questions to support learning by doing in textbooks at scale. Application of artificial intelligence in accordance with learning science research has significant potential for benefiting students.

This large-scale analysis of automatically generated questions included almost a million unique questions, more than three hundred thousand students, and more than seven million total question attempts. The substantive volume of data collected offers a distinctive perspective on not only the performance of these AG question types but also student behavior patterns. The difficulty and persistence performance metrics were qualitatively consistent with prior research on AG questions within courseware—the recognition question types were less difficult and had higher persistence than the recall question types. The difficulty results for the AG matching and FITB questions were slightly lower but within close range of previous findings—an encouraging result given the difference in learning contexts for this larger aggregated data set compared to the previous research conducted within specific university courses. However, when focusing on a subset of the current data from the use of

these questions as assignments in a university course, both the difficulty index and persistence rate increased to levels comparable to the courses in prior research. This suggests that the implementation of these questions within a classroom learning context influences how students interact with them.

This data set also enabled an exploration of how students chose to interact with text entry questions, whether FITB or the free response metacognitive tutorials. Given the very large data set, it is plausible that student effort levels would vary. Analysis of student answers revealed that a small proportion of students input non-genuine answers for the FITB questions (12.2%), though many of these students persisted to input the correct response (46.5%). By comparison, in a classroom setting the percentage of non-genuine responses was higher (17.1%) but so was persistence (92.0%). The free response questions had a diverse range of response lengths, but the majority of students made a reasonable attempt, incorporating one or both of the key terms in the tutorial activity in their answer. There was also a relationship observed between the number of key terms used and the length of the answer.

This study also suggests new avenues for future research. Additional research questions that emerge include: How does engaging in this learning by doing behavior during reading impact student behavior in their learning environments (e.g., class participation or learning outcomes)? Are certain patterns of behavior more beneficial for learning than others? How can this data be employed to refine the questions generated or the user interface? Though the success of automatically generated questions at scale is becoming established, further optimization of this learning tool will only continue to benefit students.

## 5. References


[1] J. Dunlosky, K. Rawson, E. Marsh, M. Nathan, and D. Willingham, "Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology," Psychological Science in the Public Interest, vol. 14, no. 1, pp. 4–58, 2013, https://doi.org/10.1177/1529100612453266

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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, nos. 1&2, pp. 48–59, 2022.

[7] 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.

[8] B. Jerome, R. Van Campenhout, and B. G. Johnson, "Automatic Question Generation and the SmartStart Application," in Learning at Scale, 2021, pp. 1–9, https://doi.org/10.1145/3430895.3460878

[9] 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.

[10] 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.
[11] 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.
[12] M. T. Chi, N. de Leeuw, M. H. Chui, and C. Lavancher, "Eliciting self-explanations improves understanding," Cognitive Science, vol. 18, pp. 439–477, 1994.
[13] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404-411, 2004, https://aclanthology.org/W04-3252
[14] 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.
[15] M. Moeyaert, K. Wauters, P. Desmet, and W. Van den Noortgate, "When Easy Becomes Boring and Difficult Becomes Frustrating: Disentangling the Effects of Item Difficulty Level and Person Proficiency on Learning and Motivation," Systems, vol. 4, no. 1, p. 14, 2016, https://doi.org/10.3390/systems4010014
[16] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," Educational Psychologist, vol. 46, no. 4, pp. 197-221, 2011, https://doi.org/10.1080/00461520.2011.611369.
[17] S. L. Mallicoat, Women, Gender, and Crime: Core Concepts, 1st ed. Thousand Oaks, CA: SAGE Publications, 2019.