Transforming Textbooks into Learning by Doing Environments: An Evaluation of Textbook-Based Automatic Question Generation

Rachel Van Campenhout[1[0000-0001-8404-6513]], Jeffrey S. Dittel[2], Bill Jerome[1], and Benny G. Johnson[1[0000-0003-4267-9608]]

[1] VitalSource Technologies, Pittsburgh, PA 15218, USA
[2] Britlan, Ltd., Oconomowoc, WI 53066, USA
rachel.vancampenhout@vitalsource.com

Abstract. Textbooks have been the traditional method of providing learning content to students for decades, and therefore have become the standard in high-quality content. Yet the static textbook format is unable to take advantage of the cognitive and learning science research on effective interactive learning methods. This gap between quality content and highly efficient methods of learning can be closed with advances in artificial intelligence. This paper will contextualize the need for improving textbooks as a learning resource using research-based cognitive and learning science methods, and describe a process by which artificial intelligence transforms textbooks into more effective online learning environments. The goal of this paper is to evaluate textbook-based automatic question generation using student data from a variety of natural learning environments. We believe this analysis, based on 786,242 total observations of student-question interactions, is the largest evaluation of automatically generated questions using performance metrics and student data from natural learning contexts known to date, and will provide valuable insights into how automatic question generation can continue to enhance content. The implications for this integration of textbook content and learning science for effective learning at scale will be discussed.

Keywords: Automatic Question Generation, Artificial Intelligence, Textbooks, Formative Practice, Learning by Doing, Doer Effect, Courseware.

1 Introduction

Textbooks are the de facto standard in quality educational content, and yet are not the standard in effective learning. Students encounter long sections of content that they must find a way to absorb, and risk reading passively with little retention. Entire disciplines of study have arisen from the need to identify techniques that will help students

learn this content (see [8]). Instructors may assign reading with the expectation that said readings will be absorbed prior to a learning activity, and yet have no way of monitoring student progress in real time. Students often do not read the textbook as instructors intend, limiting learning gains [9, 19]. This disconnect between textbook content and learning produces a tension that ultimately puts students at a disadvantage.

Research in cognitive and learning science has proven that certain methods are markedly more effective for learning content in online contexts. For example, integrating formative practice questions with short sections of content creates a method of learning by doing that has been shown to increase student learning gains while also decreasing the amount of time students spend studying [16]. Doing practice while reading had a six times larger relationship to learning outcomes than just reading the content [12], and follow-up research has shown this "doer effect" to be causal [13, 14, 17]. Doing practice is widely understood to be beneficial for learning, but learning science has identified that doing practice at frequent intervals while reading causes learning.

Given this simple yet highly effective learning method, why has it not been incorporated into every online textbook? The main contributing factor is an issue of scale. The volume of formative practice items needed to engage students in this type of learning by doing is in the hundreds or thousands for a typical full-semester course—a scale that becomes prohibitive in both time and cost. Question writing is a labor-intensive process that requires both subject matter and item writing expertise. The formative practice element of the courseware learning environment is therefore often a barrier too high for either content providers or teaching faculty to overcome.

Artificial intelligence is a promising solution for overcoming this barrier. Recent advances in natural language processing (NLP) and machine learning (ML) have provided tools needed to take high-quality textbook content and transform it to courseware—a process that organizes content into shorter, topical lessons and creates and embeds formative practice within those lessons. Automatic question generation (AQG) has gained increasing focus in recent years, and yet few studies have evaluated these questions empirically in a natural learning context using student data [15]. This paper makes the following contributions to the AQG literature beyond the recent systematic review by Kurdi et al. [15]. Data are evaluated from 945 students across six textbook-based courseware environments containing 2,610 automatically generated (AG) questions, making this the largest study on natural student usage of AG questions reported as of that review. We evaluate AG questions alongside human-authored (HA) questions in the same courseware on three key performance metrics: engagement, difficulty, and persistence. Our initial smaller-scale research found AG and HA questions to be similar on these metrics, suggesting students did not perceive a difference that caused them to behave differently with the AG questions [22]. Prior studies have focused on difficulty but not on engagement-based metrics crucial to formative practice.

2 Methods

2.1 The SmartStart Process

VitalSource is engaged in a large-scale project, called SmartStart [7], for automatically creating learning by doing courseware from textbooks. SmartStart uses NLP and ML

methods to accomplish three main tasks: identify short lessons of content to appear on a courseware page, find and align learning objectives to those lessons of content, and generate formative practice questions for each lesson page. These actions are completed in an application interface with options for customization by a course designer. The goal is to use this process to easily and quickly create foundational courseware that engages students in learning by doing as they read the traditional textbook material. The courseware that this process produces can be used as is, or can be further customized in an authoring interface to include adaptivity and additional assessments.

Content Sections. Most often, textbooks contain units or chapters on broad topics that cover significant amounts of content in contiguous blocks. While there may be a series of subheadings, the content flows together in a single unit. This large volume of reading poses a long-held concern about students becoming passive readers and having difficulty reading for understanding [3]. A set of expert rules for the content chunking process was derived from prior experience with several dozen custom-built courseware developed with the methods of Carnegie Mellon University's Open Learning Initiative [16]. These environments were used by thousands of students and historical data provided insights on how different lesson lengths affected student retention and activity. Using these rules, SmartStart analyzes the textbook structure and proposes how the content could be presented in shorter, topically aligned lessons in a courseware interface. It also uses expert rules to identify any potential issues with the content sections for human evaluation.

Learning Objectives. In addition to shorter sections of content, learning objectives help students construct mental models by providing clear guidance on what they are expected to learn, as well as an indication as to how they will be evaluated on that content. Learning objectives also provide a practical function in courseware environments, as they are tagged to formative practice and feed data to instructor dashboards that are organized by objective in addition to post hoc analyses. Most textbooks provide students with these types of learning statements, but their phrasing and location are highly variable. Therefore, SmartStart must first locate the learning objectives. The lack of consistency in phrasing, format, placement, and HTML markup observed across even just a few dozen textbooks made a rule-based system for locating them infeasible, so instead, a supervised ML model is used. The model includes features that represent specific identifiers, placement characteristics, and Bloom's Taxonomy verbs [4]. The learning objective identification model and other ML models used in SmartStart were developed using the scikit-learn library [18]. Once the learning objectives have been located and extracted, the next task is to place them with the lessons created in the structure task. Another model evaluates the content of each learning objective and lesson and proposes the best placement. Placing the learning objective with lesson content will also tag any formative questions placed on that lesson page, completing the data collection and instrumentation architecture of the courseware.

Automatic Question Generation. The key feature that turns static textbook content into a learning by doing environment that takes advantage of the doer effect is formative practice questions. Two types of formative cloze questions are generated by this AQG process: fill-in-the-blank (FITB) and matching (as seen in Figure 1). The FITB provides

a sentence with a term missing for students to enter, making this question a recall type on Bloom's cognitive process dimension [4]. The matching question provides a sentence with three missing terms and the student must drag and drop the terms into the correct location, making this a recognition type on that dimension. Both recognition and recall questions have long been researched for their learning value [5] and are both on the first level of Bloom's Taxonomy [4]. As formative practice, students can continue to answer until they reach the correct response and receive immediate feedback.

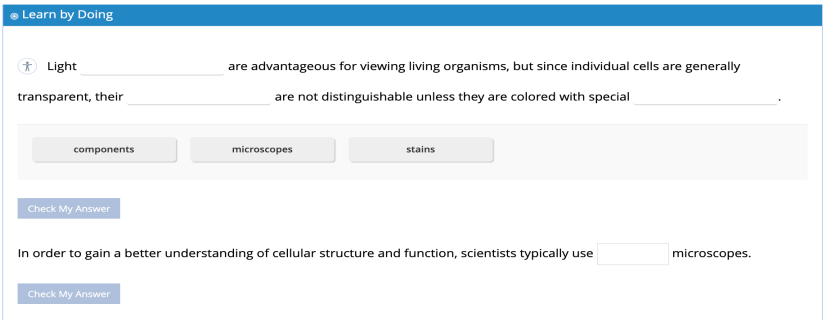

Fig. 1. Example of AG matching and FITB questions from the Microbiology courseware.

AQG is the most complex step in the SmartStart process given the number of requirements and variables involved. AQG has been an increasingly researched topic given its relevance to several fields, and there have been many differing approaches developed. To describe the current approach, we will use the classification system developed in [15]: level(s) of understanding and procedure(s) of transformation. The level of understanding for this AQG uses both syntactic and semantic information from the textbook. The NLP analyses are carried out using the spaCy library [11]. This information is used to accomplish two primary tasks: selecting the content sentences for the questions and selecting the term(s) to be used as the answer(s). Syntactic information, such as part-of-speech tagging and dependency parsing, is used in both sentence selection and answer term selection. Semantic knowledge is also used for detecting important content. The procedure of transformation is primarily a rule-based method. A set of rules is used to select the question sentences and answer terms, and these rules use both syntactic and semantic information to select the best options.

After the syntactic and semantic processing of the textbook has been completed and the selection rules applied, a set of questions has been generated. However, this set contains more questions than will ultimately appear to students. The SmartStart AQG uses an overgenerate-and-rank approach [10] to select only the top questions of each type to appear on the lesson page. As these questions are to be used by students in their natural learning environment and not as part of an experiment, the question sets were further scrutinized in a human review pass. The goal of this review was not to evaluate the questions from the perspective of a subject matter expert, but rather to search for quality issues common to the field of AQG. For example, some questions may be too easily guessable (e.g., "The father of psychoanalysis was Sigmund _____." or have grammatical problems such as an unresolved anaphoric reference in the question stem. For two of the six courses in this study (Communication A and Accounting, Table 1),

the textbook's publisher also did a subsequent human review pass and made additional minor modifications to some of the remaining questions.

2.2 Question Evaluation

The purpose of this work is to expand upon the current AQG literature by furthering the empirical evaluation of AG questions. As noted in Kurdi et al.'s 2020 systematic review [15], the majority of studies generate questions for experimental settings. Only one study reported using AG questions in a class setting [27], but this study used template-based programming exercises and evaluated student pre- and post-test scores, not individual item characteristics. Furthermore, only 14 of the 93 studies that met the criteria for the review evaluated question difficulty. These studies primarily used small samples (under 100 questions) and evaluation was based on expert review, not natural student data (see [15]). This type of expert review was critical to the development of the current AQG system, but the true test is how the AG questions perform with students in authentic learning contexts.

These AG questions were evaluated using data from a set of six SmartStart courses that were created from existing textbooks and used by students in their natural learning environments. For example, an introductory behavioral neuroscience textbook [26] was used to generate the courseware discussed in detail in the next section. These courses were also enhanced with human-authored questions post-generation, providing a unique opportunity to compare AG and HA questions that the same students completed on the same lesson pages; details are given in Table 1. The manually added HA questions can also be categorized as recognition or recall. The recall category includes the AG FITB as well as the HA FITB (the most direct counterpart) and, in the Neuroscience course, HA numeric input. The recognition category includes the AG matching and all other HA question types. Most similar to the AG matching are the HA drag-and-drop (D&D) types and the pulldown type, where students select a term from a dropdown menu to complete a question stem. There are three types of HA multiple-choice variations: conventional multiple-choice (MC), multiple-choice multiple-select (MCMS), and multiple-choice grid (MC grid). The Neuroscience course also contains HA passage selection questions, in which students select content in a short passage of text according to the instructions.

This type of in vivo experimentation across a variety of courses provides greater external validity, while comparing interactions of the same students with both types of question improves internal validity. To evaluate both question types through an empirical approach, the performance metrics of engagement, difficulty, and persistence provide a basis for comparison [22].

The first metric studied is engagement—whether or not students chose to answer questions they encountered on a lesson page. For questions that were answered, a difficulty metric can provide insights into whether questions may be too easy or difficult. The last metric is persistence—when students initially answer a question incorrectly, how often do they continue to answer until they reach the correct answer? While mean performance metric values are insightful, a mixed effects logistic regression model will also be used to analyze these metrics more rigorously, including controlling for covariates. The results will be presented and discussed in detail for a single course, followed by discussion of patterns observed across all courses.

Table 1. SmartStart courses with students and questions per course.

Course	Institutions	Students	AG Questions	HA Questions
Neuroscience [26]	18	516	747	888
Communication A [1]	1	109	263	390
Microbiology [21]	1	99	416	690
Psychology [6]	1	91	607	48
Communication B [2]	3	79	386	533
Accounting [20]	1	51	191	403

3 Results & Discussion

3.1 Neuroscience

Engagement. Doing the formative questions is what generates the doer effect, which helps students learn, so the first step is to evaluate how students choose to engage with the different question types. We hypothesize that if students perceived problems with the AG questions, they would engage with them less than HA questions or simply wouldn't do them at all. The Neuroscience course will be used as a detailed example. This course has the largest data set, and also has the largest variety of HA question types, allowing broader comparison with the AG questions. This textbook-based courseware was used at 18 institutions in 21 total course sections, providing a wide range of contexts and the highest likelihood of heterogeneity of students. The data set was constructed as *the set of all opportunities* a student had to engage with a question. While at first a simple cross of all students with all questions would seem appropriate, there are many cases where a student did not visit a lesson page, and therefore did not have the opportunity to choose whether to answer those questions. Rather, engagement opportunities were taken as all student-question pairs on pages that the student visited (very short page visits of under 5 seconds were excluded).

Given this data set of student engagement opportunities, why not just use mean engagement to assess the question types? Data from courseware show that engagement typically declines over the course of a semester, and even within modules/chapters and pages [23]. The location of a question within the course may therefore impact the likelihood that students will engage with it, and so a more sophisticated model is needed to take this into consideration. Logistic regression can be used to model the probability that a student will answer a question they encountered as a function of question type, while also taking question location variables into account as covariates. Furthermore, because there are multiple observations per student and question, these are not independent, and a mixed effects model is required. The AG FITB is the baseline for the question type categorical variable, facilitating comparison between this AG recall type and the other AG and HA types. The R formula that expresses the model is:

```
glmer(answered ~ course_page_number + module_page_number
               + page_question_number + question_type
               + (1|student) + (1|question),
               family=binomial(link=logit), data=df)
```

The data set for the Neuroscience course consists of 286,129 individual student-question observations. An answered question was recorded as 1 and an unanswered question as 0. If we first consider the mean engagement of each question type, the range is from 43.4% to 29.7%. There are clusters of question types with similar means. For instance, AG matching, HA MC, HA MCMS, and HA numeric input all had engagement around 43%. Next, AG FITB (41.1%), HA pulldown, and HA D&D table all had between 40% and 42% engagement. HA D&D image, HA FITB, and HA passage selection all had engagement below 38%. This information on its own is useful; however, reviewing the results of the model in Table 2 gives additional insights.

Table 2. Engagement regression results for the Neuroscience course.
Significance codes: 0 *** 0.001 ** 0.01 * 0.05

Fixed Effects	Mean	Significance	Estimate	*p*
Intercept		***	-2.17527	< 2e-16
Course Page		***	-0.74925	< 2e-16
Module Page		***	-0.31960	< 2e-16
Page Question		***	-0.09011	9.37e-06
HA D&D Image	29.7		-0.19026	0.700107
HA D&D Table	41.7		0.27267	0.356745
HA Pulldown	40.2	**	0.20531	0.009303
AG Matching	43.3	***	0.22083	0.000497
HA MC	43.4	***	0.24570	0.000536
HA MCMS	43.2	*	0.19886	0.017558
HA Passage Selection	30.4	***	-1.52872	0.000879
HA FITB	37.1	**	-0.21440	0.004556
HA Numeric Input	43.3		-0.13641	0.421057

The variables for the location of the questions were all significant ($p < 0.001$) and negative, verifying that students are less likely to engage with the practice as they get to the end of a page, module, and course. After controlling for the effects of question location, we can examine differences in engagement for the question types. The HA MC and AG matching questions, which had nearly identical mean engagement scores, also had very similar estimates and significance ($p < 0.001$) compared to the AG FITB. Both question types are recognition types, but one was human-authored and one automatically generated. The HA pulldown and HA MCMS were both similar in mean scores and both more likely to be answered ($p < 0.01$ and $p < 0.05$ respectively). The AG FITB also had similar mean engagement with the HA numeric input, and the model showed no significant difference for engagement between these recall question types. Neither of the D&D question types were significantly different from the AG FITB, despite one having similar mean engagement scores and the other having much lower mean scores. The HA passage selection had the lowest mean engagement, and also was significantly less likely to be engaged with ($p < 0.001$). This could be due to the complexity of the question type—requiring students to interpret instructions, read content, and then select a segment of that content. Finally, the HA FITB (as the most direct comparison to the AG FITB) had a lower mean score and the model showed that students were less likely to engage with this type than the AG FITB ($p < 0.01$).

Difficulty. When students choose to answer the formative questions, we can evaluate the question difficulty through their first attempt accuracy. This difficulty data set consists of all attempted questions for a total of 120,098 observations, with a correct answer recorded as 1 and an incorrect answer recorded as 0. When we first consider the mean difficulty scores per question type (Table 3), we see that there is a very wide range, from a high of 86.4% correct for HA D&D image to a low of 33.7% for HA passage selection. There are three general groupings of question types. The easiest questions, with means above 80%, were the HA D&D table, AG matching, and HA D&D image. The next group ranges from a high of 70.0% to a low of 64.1% (AG FITB) and includes the HA recognition types pulldown and MC, as well as all HA and AG recall types (HA FITB, HA numeric input, AG FITB). The most difficult question types were HA MCMS at 43.3% and HA passage selection at 33.7%.

The same mixed effects logistic regression model was used, but with the difficulty data set (Table 3). The location variables, unlike for engagement, were not all significant. The location of the question in the module was significant ($p < 0.01$), but neither the course page nor the location on the page was significant. Unlike for engagement, there was no consistent trend of location significance for difficulty or persistence, and so for brevity the location variable results are omitted from the remaining tables (full results are available at [25]), though the question type regression results are still controlled for location. The regression model results for the question types were generally consistent with the trends for the unadjusted mean scores. Nearly all the questions with higher mean scores than the AG FITB (64.1%) were also easier than the AG FITB with varying degrees of significance. This also includes the HA FITB—the most direct counterpart to the AG FITB. The two question types with lower means—HA passage selection and HA MCMS—were also more difficult than the AG FITB in the model. Interestingly, the question type that was not statistically different from the AG FITB was another recall type, HA numeric input.

Table 3. Difficulty regression results for the Neuroscience course

Fixed Effects	Mean	Significance	Estimate	*p*
HA D&D Image	86.4	*	1.47548	0.041173
HA D&D Table	80.8	*	1.09198	0.011490
HA Pulldown	70.0	***	0.44359	0.000107
AG Matching	84.3	***	1.44140	< 2e-16
HA MC	67.8	**	0.27696	0.007248
HA MCMS	43.3	***	-1.06100	< 2e-16
HA Passage Selection	33.7	*	-1.52609	0.025964
HA FITB	69.0	*	0.26882	0.014033
HA Numeric Input	68.6		0.31414	0.213834

Persistence. The last performance metric is persistence—when a student gets a question incorrect on their first attempt, do they continue to answer until they reach the correct response? The persistence data set is therefore a subset of the difficulty data set, including only the incorrect first attempts for questions for a total of 34,124 observations. If a student eventually achieves the correct response the outcome was recorded as 1, and if they did not persist until reaching the correct response the outcome was recorded as 0. We hypothesize that persistence could be related to the difficulty of a

question type; if a student perceives a question type as too easy or too difficult, they may not persist until reaching the correct response as often. It may seem that this persistence metric is less impactful than engagement or difficulty, but VanLehn notes in a meta-analysis of literature on human and computer-based tutoring that getting students to finish problems correctly instead of giving up has a strong impact on learning [24]. The persistence metric should therefore not be overlooked.

The table for persistence was omitted for brevity (see [25]), as the trends were less complex. There are six question types between 98.9% and 100% persistence, and all are recognition types, including the AG matching. The next group is the three recall question types: AG FITB (89.0%), HA FITB (85.7%), and HA numeric input (80.8%). The outlier is the HA passage selection at 42.3%. Comparatively, the regression model provides interesting results. Only three question types are significantly more likely ($p < 0.001$) to be answered until correct compared to the AG FITB: HA pulldown, AG matching, and HA MC. Note that these question types were statistically easier than the AG FITB, but they ranged from 67.8% to 84.3% correct. All three are recognition type questions, which may account for the range in difficulty yet similarities in persistence.

The HA MCMS was an outlier in terms of its high degree of difficulty, and yet had statistically higher persistence than the AG FITB. There were also three question types with significantly lower likelihood of persistence. The HA passage selection is the most expected, given its high difficulty and low mean persistence. However, the other two HA recall types had statistically lower persistence than the AG FITB. Students were more likely to persist in the AG FITB than either the HA FITB or HA numeric input.

3.2 Trends Across Courses

As seen in the regression models for each of the performance metrics above, there are valuable insights into how students engage with different question types. In this section, we present the mean values with the effects from the regression model for each course, for each metric. Trends as well as anomalies can be detected when looking across courses that can provide a higher degree of generalizability for findings and suggest interesting areas for future research. Each course was different in terms of the HA question types that were added to the SmartStart courseware. To discern trends that may be characteristic of a question type, only types appearing in at least three of the six courses are presented in the tables below; however, all of each course's question types were included in its regression model (see [25]). If a question type was not used in a course, its cells contain n/a. The mean is presented for every question type, but only significant effects from the model are presented with their signs for ease of interpretation.

Engagement for All Courses. First, the mean engagement trends for question types (Table 4) show similar patterns to the Neuroscience course. The AG matching questions have means within a few percentage points of the HA MC for all courses, and often close to other types such as the HA pulldown or HA MCMS. This shows that engagement is similar for all these recognition types, regardless of the AG or HA origins. Similarly, the mean engagement for the AG FITB is typically within a few percentage points of the HA FITB, with three of four courses showing the AG FITB with slightly higher engagement, indicating the recall question types are generally close in mean engagement.

The mean engagement values also show variation between courses in their overall spread. In some courses, mean engagement for all question types is very close, such as for Communication A (43.4–50.5%), Communication B (44.8–53.3%), or Psychology (84.9–87.4%). Other courses have a wider variation in question type engagement means, such as Accounting (52.0–77.9%) or Microbiology (42.3–69.3%). Courses also have different ranges from one another; while Communication A and B have low mean engagement across all question types, Psychology has much higher mean engagement for all question types. While the reason for these differences is not discernible from the data, it is likely that the implementation of the courseware in the classroom could be a strong contributing factor, as data have shown that instructor implementation practices can greatly influence student engagement with formative practice [23].

Considering the effects from the regression model, none of the question types had consistent positive or negative significant differences in engagement compared to the AG FITB across all courses. The AG matching was positively significant in five out of six courses, and the HA MC had positive significant engagement in four out of six courses. The HA pulldown was positively significant in four courses and negatively significant in one. Interestingly, the HA MCMS was not significantly different from the AG FITB in three of five courses. While this question type is a recognition type, it is also more complex than others like the HA MC or pulldown, which could be impacting how often students choose to engage. The HA FITB was not significantly different from AG FITB in three courses, and negatively significant in one. As the recall counterpart to AG FITB, this strongly suggests that students treat recall questions similarly regarding engagement. These trends indicate that the context of the course implementation will likely influence the overall engagement patterns, but there is not evidence to suggest that students engaged with AG question types differently than similar HA types.

Table 4. Engagement means and regression model effects for all courses.

	Neurosci		Comm A		Microbio		Psychology		Comm B		Accounting	
Observations	286,129		43,538		71,119		34,757		35,351		14,661	
HA D&D Table	41.7		50.5	+**	47.1		n/a	n/a	n/a	n/a	52.7	
HA Pulldown	40.2	+**	48.5	+*	42.3	–***	n/a	n/a	53.3	+***	72.0	+**
AG Matching	43.3	+***	45.1	+*	66.9	+***	85.0	+***	52.1	+***	57.6	
HA MC	43.4	+***	43.4	+*	69.3		87.4	+***	49.9	+***	52.0	
HA MCMS	43.2	+*	47.1		54.9		n/a	n/a	51.7	+***	77.9	
AG FITB	41.1	n/a	49.5	n/a	61.9	n/a	84.9	n/a	47.2	n/a	62.5	n/a
HA FITB	43.3		45.0		n/a	n/a	n/a	n/a	44.8		58.9	–***

Difficulty for All Courses. The difficulty means show trends consistent across courses (Table 5). For instance, each course shows a range of difficulties across question types, generally from the mid-forty to eighty percent accuracy range. This aligns with expectations that some question types may be easier or more difficult than others.

The effects from the regression model show differences in question difficulty when we compare question types to the AG FITB. The AG matching question type is significantly easier ($p < 0.001$) than the AG FITB in every course. The HA pulldown is also statistically easier in four courses, with no statistical difference in one course. The HA

MC questions present an interesting mix across courses. They are statistically easier than the AG FITB in three courses, not statistically different in one course, and statistically more difficult in two courses. While the HA MC is a recognition type, which typically trends easier than recall, the question type does not solely determine the difficulty level. In particular, a question's content obviously also contributes to its difficulty; one possible explanation for the observed results is that the MC format has more flexibility for question authoring than certain other formats, including FITB, as the entire text of the stem and distractors are up to the author. For example, it is more feasible to create higher-level Bloom's questions in a MC format than one where the answer is a single word. Although beyond the scope of the present study, this is an interesting area for future investigation. Similarly, the HA MCMS is a recognition type but is statistically more difficult than AG FITB in three courses and not different in two courses.

For the recall question types, the HA FITB was statistically easier than the AG FITB in one course and more difficult in one. For the remaining two courses, the HA FITB was negative with marginal significance ($p = 0.0521$ and $p = 0.06233$); while not reported as significant, this points to an interesting trend. Given the similarity of these question types, differences in difficulty should be investigated further in future work.

Table 5. Difficulty means and regression model effects for all courses.

	Neurosci		Comm A		Microbio		Psychology		Comm B		Accounting	
Observations	120,098		20,990		42,114		29,583		17,547		8,309	
HA D&D Table	80.8	+*	75.6		56.7		n/a	n/a	n/a	n/a	84.2	+***
HA Pulldown	70.0	+***	71.7		68.2	+**	n/a	n/a	79.8	+***	86.7	+***
AG Matching	84.3	+***	81.6	+***	90.5	+***	89.8	+***	81.0	+***	86.5	+***
HA MC	67.8	+**	65.8	–*	69.6	+***	64.6	–***	75.5	+***	63.3	
HA MCMS	43.3	–***	57.6		43.3	–**	n/a	n/a	50.7	–***	54.8	
AG FITB	64.1	n/a	73.1	n/a	62.8	n/a	89.9	n/a	68.6	n/a	60.1	n/a
HA FITB	69.0	+*	45.1	–***	n/a	n/a	n/a	n/a	63.6		28.4	

Persistence for All Courses. The table of persistence means and regression model effects is omitted for brevity (see [25]). Of the three metrics, persistence had the least variation across question types and courses, and so we simply summarize the results. The persistence data sets ranged from 34,124 observations (Neuroscience) to 2,027 observations (Accounting). Mean persistence is generally high across all question types and courses. High persistence is ideal, as this means students continue to answer until they reach the correct response. Mean persistence was over 80% in all but a few instances. Several recognition types had consistently high persistence in all courses: HA pulldown, AG matching, and HA MC. These also all have positive significance compared to the AG FITB, with the exception of one HA MC case that was not significantly different. These are encouraging findings, as the AG matching are grouped very closely with other HA recognition types. The AG matching was generally one of the least difficult question types, and yet that did not discourage students from persisting when they did answer them incorrectly. The HA FITB had negative significant results ($p < 0.001$) compared to the AG FITB in each course. Students were less likely to persist on the

HA FITB, which is interesting given how similar the question type is to the AG FITB. This trend will be a focus of future research.

4 Conclusion

The current advances in artificial intelligence, natural language processing, and machine learning make it possible to take high-quality textbook content and automatically transform it into learning by doing courseware designed to be more effective for student learning. AQG that is directly based on the textbook content is a practical solution to achieve this goal of combining content and learning science-based practices at large scale. However, the AG questions should be rigorously evaluated to ensure they meet certain standards, and a comparison of performance metrics to HA questions provides the first step to ensuring the quality of these questions.

Through the analysis of data from 945 students who used six textbook-based courseware containing a total of 2,610 AG questions, several interesting trends were revealed. Student engagement with AG matching questions was very similar to other HA types such as multiple choice or pulldown. The AG FITB was similar in engagement to the HA FITB and numeric input. This indicates that there was not a difference in engagement between comparable AG and HA question types, but rather that the recognition vs. recall nature of the question type had the greatest impact. When considering difficulty, the AG matching was generally one of the easiest question types while the AG FITB was typically in the middle of the range. Across courses it was seen that the difficulty of question types can vary. The HA MC questions were sometimes easier and sometimes more difficult than the AG FITB, suggesting the influence that content can have on the difficulty of question types. Finally, we see high persistence rates for most question types in general. The AG matching was very similar in persistence to other HA recognition types, indicating the easier questions did not deter students from answering them until correct. However, the HA FITB—sometimes more difficult than the AG FITB—had statistically lower persistence than its AG recall counterpart.

Overall, the trends from this large-scale data analysis indicate that students in natural learning contexts do not treat the automatically generated questions differently than their human-authored counterparts. Along with continuing to expand the scope of courses analyzed, future research will involve comparison of AG and HA questions on additional metrics such as discrimination and alignment. As previously noted, some of the unanticipated results also suggest interesting avenues for investigation. The ultimate validation, however, will be to investigate the impact of AG questions on student learning directly, such as in replication of studies of the doer effect on summative assessments [12–14, 17]. There is still much more to learn, but these findings give optimism that textbook-based, automatically generated questions could provide a scalable path to delivering the learning benefits that have been shown with human-authored questions. This will continue to be a major research focus for some time to come.

Acknowledgment
The authors gratefully acknowledge Cathleen Profitko for identifying courses for these analyses, Nick Brown for assistance with [25], and the anonymous reviewers for their thoughtful and constructive comments.

References

1. Adler, R. B., Rodman, G., & du Pré, A. (2019). Essential Communication (2nd ed.). New York: Oxford University Press.
2. Adler, R. B., Rosenfeld, L. B., & Proctor II, R. F. (2021). Interplay: The process of interpersonal communication (15th ed.). New York: Oxford University Press.
3. Adler, M. J., & Van Doren, C. (1940). How to read a book. New York, NY: Touchstone/Simon & Schuster.
4. Anderson, L. W. (Ed.), Krathwohl, D. R. (Ed.), Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives (Complete edition). New York: Longman.
5. Andrew, D. M., & Bird, C. (1938). A comparison of two new-type questions: recall and recognition. Journal of Educational Psychology, 29(3), 175–193. https://doi.org/10.1037/h0062394
6. Bosson, J. K., Vendello, J. A., & Buckner, C. V. (2018). The psychology of sex and gender (1st ed.). Thousand Oaks, California: SAGE Publications.
7. Dittel, J. S., Jerome, B., Brown, N., Benton, R., Van Campenhout, R., Kimball, M. M., Profitko, C., & Johnson, B. G. (2019). SmartStart: Artificial Intelligence Technology for Automated Textbook-to-Courseware Transformation, Version 1.0. Raleigh, NC: VitalSource Technologies.
8. Dunlosky, J., Rawson, K., Marsh, E., Nathan, M., & Willingham, D. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. Psychological Science in the Public Interest, 14(1), 4–58. https://doi.org/10.1177/1529100612453266
9. Fitzpatrick, L., & McConnell, C. (2008). Student reading strategies and textbook use: an inquiry into economics and accounting courses. Research in Higher Education Journal, 1–10.
10. Heilman, M., & Smith, N. A. (2009). Question Generation via Overgenerating Transformations and Ranking. Retrieved February 9, 2021, from www.lti.cs.cmu.edu
11. Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303
12. Koedinger, K., Kim, J., Jia, J., McLaughlin, E., & Bier, N. (2015). Learning is not a spectator sport: doing is better than watching for learning from a MOOC. In: Learning at Scale, pp. 111–120. Vancouver, Canada. http://dx.doi.org/10.1145/2724660.2724681
13. 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. Learning Analytics and Knowledge. Edinburgh, United Kingdom. http://dx.doi.org/10.1145/2883851.2883957
14. Koedinger, K. R., Scheines, R., & Schaldenbrand, P. (2018). Is the doer effect robust across multiple data sets? Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018, 369–375.
15. 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(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y
16. Lovett, M., Meyer, O., & Thille, C. (2008). The Open Learning Initiative: Measuring the Effectiveness of the OLI Statistics Course in Accelerating Student Learning. Journal of Interactive Media in Education. http://doi.org/10.5334/2008-14

17. Olsen, J., & Johnson, B. G. (2019). Deeper collaborations: a finding that may have gone unnoticed. Paper Presented at the IMS Global Learning Impact Leadership Institute, San Diego, CA.
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. In Journal of Machine Learning Research (Vol. 12). http://scikit-learn.sourceforge.net
19. Phillips, B. J., & Phillips, F. (2007). Sink or Skim: Textbook Reading Behaviors of Introductory Accounting Students. Issues in Accounting Education, 22(1), 21–44. https://doi.org/10.2308/iace.2007.22.1.21
20. Scott, P. (2019). Accounting for business (3rd ed.). New York: Oxford University Press.
21. Swanson, M., Reguera, G., Schaechter, M., & Neidhardt, F. (2016). Microbe (2nd ed.). Washington, DC: ASM Press
22. Van Campenhout, R., Brown, N., Jerome, B., Dittel, J. S., & Johnson, B. G. (2021). Toward Effective Courseware at Scale: Investigating Automatically Generated Questions as Formative Practice. Learning at Scale. https://doi.org/10.1145/3430895.3460162
23. Van Campenhout, R. & Kimball, M. (2021). 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.
24. VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369
25. VitalSource Supplemental Data Repository. https://github.com/vitalsource/data
26. Watson, N. V., & Breedlove, S. M. (2021). The mind's machine: Foundations of brain and behavior (4th ed.). Sunderland, Massachusetts: Sinauer Associates.
27. Zavala, L., & Mendoza, B. (2018). On the use of semantic-based AIG to automatically generate programming exercises. In: the 49th ACM technical symposium on computer science education, ACM, pp. 14–19.