The Content Improvement Service: An Adaptive System for Continuous Improvement at Scale

Bill Jerome, Rachel Van Campenhout(✉), Jeffrey S. Dittel, Richard Benton, Scott Greenberg, and Benny G. Johnson

VitalSource Technologies, Pittsburgh, PA 15218, USA
{bill.jerome,rachel.vancampenhout}@vitalsource.com

Abstract. Advances in artificial intelligence and automatic question generation have made it possible to create millions of questions to apply an evidence-based learn by doing method to thousands of e-textbooks, an unprecedented scale. Yet the scaling of this learning method presents a new challenge: how to monitor the quality of these automatically generated questions and take action as needed when human review is not feasible. To address this issue, an adaptive system called the Content Improvement Service was developed to become an automated part of the platform architecture. Rather than adapting content or a learning path based on student mastery, this adaptive system uses student data to evaluate question quality to optimize the learning environment in real time. In this paper, we will address the theoretical context for a platform-level adaptive system, describe the methods by which the Content Improvement Service functions, and provide examples of questions identified and removed through these methods. Future research applications are also discussed.

Keywords: Content Improvement Service · Adaptive instructional systems · Iterative improvement · Grey-box systems · Artificial intelligence · Automatic question generation

1 Introduction

Adaptive instructional systems have varied widely in the past decades. Different technologies have focused on a range of adaptive strategies—including content level of difficulty, tutoring dialogues, self-correction, metacognitive prompts, etc.—with varying levels of effectiveness [2]. Systematic reviews have worked to make sense of a diverse field of research through categorizations and frameworks. Vandewaetere et al. [14] describes adaptive systems according to the source of adaptation (what determines adaptation), target of adaptation (what is being adapted), and pathway of adaptation (how it is adapted). Vandewaetere et al. identify the primary source of adaptation as learner characteristics that point to aptitude characteristics or a learner model. Martin et al. [8] expanded the adaptive source to include a content model and instructional model in addition to the learner model. Yet no matter the source of adaptation, we can see that these systems generally focus on an evolving relationship between the learner and the

G. Meiselwitz et al. (Eds.): HCII 2022, LNCS 13517, pp. 286–296, 2022.
https://doi.org/10.1007/978-3-031-22131-6_22

pedagogical path. What if the source and target of adaptation are one and the same? This paper will outline an adaptive instructional system which does not focus on an individual student as the source of adaptation, but rather focuses on individual question items to adaptively update the learning environment.

While Vandewaetere et al. [14] noted the increasing use of intelligent learning for adaptive systems, it was also still an evolving technology with unknown future impact. "Currently, Bayesian networks, fuzzy logic and neural networks are considered as new approaches to the development of learner models. However, based on our review, we can conclude that all of these newer techniques are still in the very early stage of development, and none of the techniques has been concretely implemented in an adaptive system," (p. 128). Artificial intelligence (AI) became part of Martin et al.'s [8] adaptive learning model framework, serving as part of the "adaptive engine" in their system (though notably this study did not address the use of AI in the literature it reviewed).

The adaptive system described herein was developed as a solution to a new challenge created by using AI for automatic question generation (AQG) on an enormous scale. Bookshelf CoachMe™ (BCM), a new learning feature of the Bookshelf e-reader platform from VitalSource Technologies, uses AQG to deliver formative practice questions alongside the e-textbook content so students can practice while they read. Millions of questions were generated using AI and released in over 4,500 e-textbooks as a free feature of the e-reader platform. The goal of these automatically generated (AG) formative questions is to help students become active participants in their learning process by practicing at the point of learning. This method of "learn by doing" has been shown to have six times the effect on learning outcomes compared to reading alone [6], and follow-up research has found that this method is causal to learning [6, 13]. As anyone who has created educational content is aware, no content is perfect. Historically, textbooks, courseware, and any learning content students see goes through review and QA prior to being released as well as after it's released, as it is inevitable that students find problems or errors. The AG questions described had extensive automated QA as well as targeted human QA prior to release. Research on questions generated through this AI process found that they performed equally as well with students as human-authored questions on several key metrics [11]. And still, as no human could write millions of perfect questions, neither will AI generate a perfect question set. So while AI has solved the problem of how to create formative practice for effective learning at scale, a new challenge presents itself in how to monitor and QA this enormous question set. The solution is an AI-driven adaptive system.

The Content Improvement Service (CIS) is an adaptive system not limited to a single instance of a course or learning environment, but rather is a platform-level system that monitors all questions delivered in all e-textbooks. The CIS uses all student responses to make decisions about the quality of the questions. In this way, the content is the source of adaptation as well as the target of adaptation. The CIS uses data at a micro level, as it is monitoring each individual question and every student answer. Yet at the same time, the CIS uses data at a macro level, as it is using millions of data points to make decisions for an entire platform.

We can begin to see the differentiation between an adaptive system that acts to move a student through a content path and one that adapts for the purpose of iterative

improvement of the content itself. The concept of iterative improvement is one familiar to many fields, but is of particular importance to learning engineering. Learning engineering is a practice and process that uses human-centered engineering design and data-informed decision-making to support learners [4]. Iterative improvement is key to the process of learning engineering, both in the development of learning experiences, as well as the action following data analysis [5]. From previous case studies, we see examples of how this data-driven iterative improvement cycle can benefit students [10], yet often the data analysis is only done infrequently at specific points due to scarcity of expert human resourcing. The CIS uses real-time question responses and takes an action to adapt the learning environment for all students. This process is not one that happens at prescribed times (such as after a semester ends) but instead it occurs continuously. The CIS performs a continuous process of iterative improvement for the AG questions across the platform, thereby automating this key learning engineering cycle.

The goal of this paper is to examine the CIS as a tool for large-scale adaptation and iterative improvement. We will describe the architecture of this system, the technology driving its function, and examine examples of the improvements it identifies for AG questions. In this way, we hope to illuminate how combining learning science-based methods with adaptive instructional systems and scalable artificial intelligence technology can produce systems of great benefit to millions of learners worldwide.

2 Architecture of the CIS

As defined by Vandewaetere et al. [14], the Content Improvement Service is the pathway of adaptation, by monitoring question data, identifying problems, and carrying out decisions. Illustrated in Fig. 1, the CIS is the critical link between the live learning platform and the passive content management system (CMS). In many online learning environments, a manual feedback loop takes place after a course is delivered in order to improve the materials for the next semester. The CIS enables us to make changes as soon as enough data are available to demonstrate a need for a change instead of waiting, e.g., for an entire semester to go by. As students enroll and proceed at different rates through an e-textbook, we have the chance to improve practice for the next student to encounter it during the same semester without the semester (or more) lag.

The CIS has two essential types of input. The first is information about the questions and content available to students in the learning platform. Knowing some basic data—such as the question identifier, the textbook it belongs to, and the type of question (multiple choice, text entry, etc.)—provides the CIS with enough basic information do its analysis. Importantly, there is no domain knowledge required. The second type of input is data about student interactions with the content and questions, for example, correct and incorrect question attempts. Additionally, students are given the option to rate questions they answer and this feedback is also considered. No Personal Identifiable Information (PII) is tracked or needed within the CIS. The continuous stream of interactions is used to update a local database with summary statistics for each question being analyzed. Although the richness of both types of data may increase as new ideas are developed for automatically improving content, the two broad categories and pipelines ultimately remain the same.

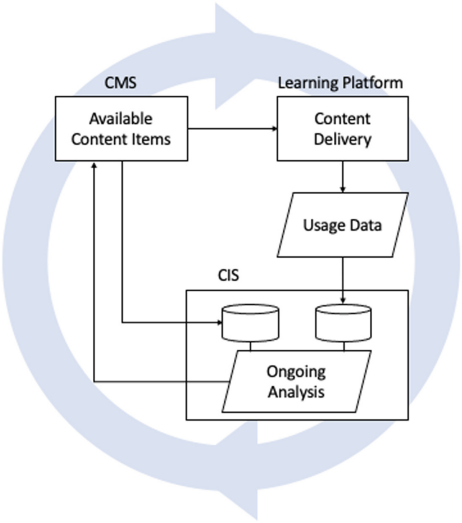

Fig. 1. An ongoing automatic cycle of content improvement based on usage data. Content is published from a CMS into a learning platform where students engage with the content. The data generated from that usage are monitored by the CIS coupled with basic information about the content it received during the publishing step. Should any rule apply to trigger content improvement, it messages the CMS with the relevant information allowing the CMS to publish the improvement to the live content in the learning platform.

In a continuous manner, the CIS updates the available summary data and determines what actions, if any, should be taken. For example, a decision rule may trigger that indicates a question is not performing acceptably and should be removed (or replaced if there is a replacement available). In response to this update from the CIS, the CMS automatically publishes an update to the live course and informs the CIS of the question removal and replacement (if any) and the cycle of automated iterative improvement continues.

3 Recall: The Guiding Philosophy of the CIS

An important design consideration for the tests of question quality is whether they should be oriented toward precision or recall. That is, is it more important to require a high degree of certainty that a question is unsatisfactory before removing it, or that as many unsatisfactory questions as possible are identified and removed? These two requirements are at odds with each other, and in general not possible to satisfy simultaneously (known as the "precision-recall tradeoff").

As a key goal of the CIS is to minimize the exposure of these questions to students, there is a need to identify them as quickly as reasonably possible. This means emphasizing recall over precision—erring on the side of caution rather than continuing to collect evidence in order to maximize confidence in the question's classification. The question generation process for BCM uses an overgenerate-and-rank approach [3] to

ensure there is a surplus of questions available for replacing questions deemed unsatisfactory, and so the system can afford to be more aggressive in removing potentially problematic questions. This focus on recall is also in service to a student-centered approach of development, another critical element of learning engineering. In this context, the precision-recall tradeoff is occurring in a student learning environment so in maximizing the recall of questions in the CIS, we are prioritizing the student experience. In this way, recall becomes a guiding philosophy of the CIS.

4 Decision-Making in the CIS

The CIS is a platform-wide adaptive system that makes decisions about questions in real-time, making it a large part of the architecture of the learning environment. Because of this, the decision-making processes are not intended to remain part of a "black-box" system where nothing is known about the system's inner workings by its users. Sharma et al. [9] put forth a rationale for using a "grey-box" approach:

> ....where the input features can be informed from the context and the theory/relevant research, the data fusion is driven by the limitations of the resources and contexts (e.g., ubiquitous, low-cost, high precision, different experimental settings), and the [machine learning] method is chosen in an informed manner, rather than just as a way to obtain the optimal prediction/classification accuracy. In other words, this contribution aims to invite researchers to shift from the optimal ends (outputs) to the optimal means (paths), (p. 3007).

In the case of the CIS, the goal of the system is to make decisions about questions as quickly as possible, and so the learning science context is critical to determining the optimal means by which the CIS makes its decisions. Designed to address the challenge of continuously evaluating the performance of formative questions at scale, the CIS was developed with a recall philosophy and a learner-centered approach. The methods for its decision-making were derived from this context; incorporating research ranging from student perception to question psychometric properties shaped the methods by which the CIS operates.

This grey-box approach also fosters trust and accountability in the system through the transparency of its research base and methods. When decisions are being made in an automated fashion that impact student learning environments at scale, the ability to explain how those decisions are made is necessary. In this paper we outline several research-based optimal means that contribute to the optimal ends—the decisions of the CIS. It is also noteworthy that CIS uses the methods selected as tools, and that these tools can be modified, extended, added or removed based on evolving technological or theoretical breakthroughs. In this way, the CIS is itself a system that can be iterated upon based on data and research.

This grey-box approach of incorporating relevant theory and research into the methods of the CIS (or pathway of adaptation) is therefore also congruent with a learning engineering approach. Research is a critical input of the learning engineering process [10], and here it is also used to determine the methods of adaptation. The CIS was developed using the learning engineering approach, and to apply this domain expertise

with the technical capacity to make research- and data-informed decisions at massive scale. It would be more difficult to trust a black-box system designed without regard for context or learning theory to make unsupervised determinations about content. Instead, a grey-box system designed by learning science experts to use research-based methods to achieve a student-centered outcome provides a responsible and accountable system.

4.1 Student Helpfulness Ratings

One simple and direct measure of question quality is student feedback, so the first method of the CIS we outline is the use of student helpfulness ratings. After answering a question in BCM, the student is given the opportunity to give it a simple thumbs up/thumbs down rating of helpfulness (Fig. 2). When a thumbs down rating is given, additional feedback can optionally be provided on why the student felt the question was not helpful.

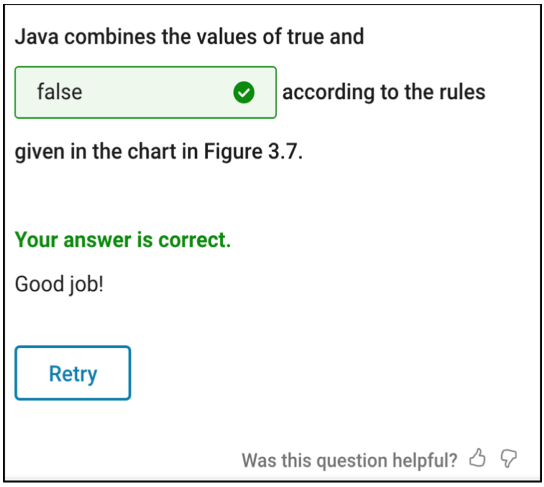

Fig. 2. Question helpfulness rating feature.

Analysis of a data set of BCM usage comprising 911,044 student-question interactions in 3,948 textbooks showed that students do not rate questions very often, only in 0.52% of rating opportunities, with a thumbs down rating 0.20% of the time.

A single thumbs down rating is generally not sufficient evidence that a question should be removed. Some students tend to challenge questions they answer incorrectly, and it is also well known that question rating agreement is often low even among trained reviewers [7], which can be common for subjective judgments. Determination of an appropriate decision rule or rules based on the number and context of thumbs down ratings requires calibration from data. However, an example from an economics textbook of a question with two thumbs down ratings shows that the additional information students can optionally provide can be helpful:

> "The _____ of a new sports car doesn't just affect the person driving off the dealer's lot."

Answer: “sale”

Both students rating the question indicated that the question was not relevant to the subject matter, which increases the likelihood that it should be removed.

4.2 Bayesian Inference of Mean Score

The CIS also considers question quality as outlined by the literature as part of its methods. Of the psychometric properties relevant to question quality, perhaps the one of most interest is difficulty. For formative practice, for example, if questions are too easy or too difficult, it may risk diminishing student engagement and satisfaction. Therefore, it is necessary for the CIS to monitor difficulty so that questions not meeting the desired criteria can be removed. A common way to gauge a question’s difficulty is by its mean score (sometimes called the difficulty index). If a question’s mean score is not acceptable, our task is to learn this as quickly as possible from the observed data (to a specified confidence level) in order to minimize exposure of poorly performing questions.

Here, a simple approach is to model each question independently, treating students’ answers to the question as Bernoulli trials. A Bernoulli random variable models an event with exactly two possible outcomes, such as success and failure (e.g., correct and incorrect), represented as 1 and 0, and has a single parameter p, the probability of success, which in this case represents the question’s mean score. The Bernoulli trials model requires that the trials are independent, and that the probability of success or failure is the same for each trial. The first is reasonably satisfied (each trial is an answer by a different student) but the second is not, since this model approximates the probability of a correct answer as the same for all students (the mean score), when in fact it depends strongly on the individual student. While this assumption would be too restrictive for many analyses, it is entirely adequate here. While a more complex model like item response theory [1]—that takes student ability into account—would be more accurate, it would also require collecting much more data to make the assessment, which is at odds with our requirements. The higher accuracy afforded is not needed when we recognize that recall is more important than precision in identifying poorly performing questions; put another way, we are perfectly willing to sacrifice some acceptable questions in order to remove the unacceptable ones. Furthermore, when the mean score of an AG question is very low it is sometimes indicative of an error in the generation process that yielded a question that is not correctly answerable, making individual student abilities less relevant.

The need to assess a question’s difficulty from a small sample of student data suggests a Bayesian approach. Bayesian methods provide a powerful and flexible approach to estimation of models from data. In particular, this enables probability distributions for a model’s parameters to be learned from data rather than simply point estimates. The Bayesian approach combines prior knowledge or assumptions about the model parameter distributions with the likelihood of the observed data under the model to obtain the joint posterior distribution of the model parameters. A Bayesian approach can thus help us arrive at better-quality decisions more quickly by allowing us to incorporate what is known about question mean scores from prior experience. The total number of successes in a given number of Bernoulli trials has a binomial likelihood function, and for Bayesian

inference it is common to use a beta distribution as the prior distribution of p since this has a closed-form solution for the posterior distribution, which is also a beta distribution.

The shape parameters α and β of the beta prior distribution can be determined by fitting the mean and variance of an empirically observed set of question mean scores. This gives a so-called "informed prior." A data set from a previous large-scale study on the difficulty of automatically generated and human-authored questions [11] was used to determine a prior in this manner. Figure 3 shows a histogram of mean scores of 809 AG questions and the beta distribution fit to it ($\alpha = 4.58$, $\beta = 1.82$).

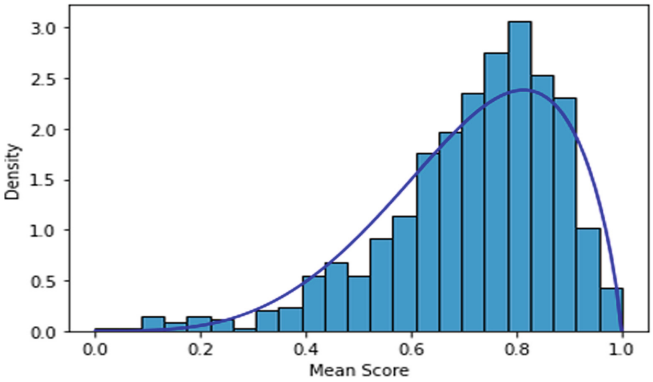

Fig. 3. Informed prior distribution for mean score obtained from AG question mean scores.

For the sake of illustration on a real example, suppose we wish to remove a question if it is at least 90% likely that its mean score is less than 0.5, i.e., more students will answer it incorrectly than correctly. A particular question in a human resources textbook had 4 correct and 16 incorrect answers in the first 20 students. Should it be removed? To decide, we must construct the posterior distribution of the question's mean score from the prior and observed data, and then evaluate the decision rule with it. The posterior is obtained by updating the prior's α and β values with the number of observed correct and incorrect answers, respectively, giving $\alpha = 8.58$, $\beta = 17.82$. The probability that the mean score is less than 0.5 is then simply the posterior's cumulative distribution function at 0.5, the shaded area in Fig. 4. Note that the posterior has been shifted significantly to the left of the prior based on the observed data. The shaded area is 0.968, or 96.8%, which is greater than the 90% threshold, so the question should be removed.

The question to be removed based on Bayesian inference on its mean score is:

> "Employees want to work for employers that can provide them with a certain amount of _____ security."
>
> Answer: "economic"

Among the incorrect answers students gave were "job," "employment," and "financial." Rather than corresponding to misconceptions, these responses are effectively synonymous with the expected answer, leading to the low mean score.

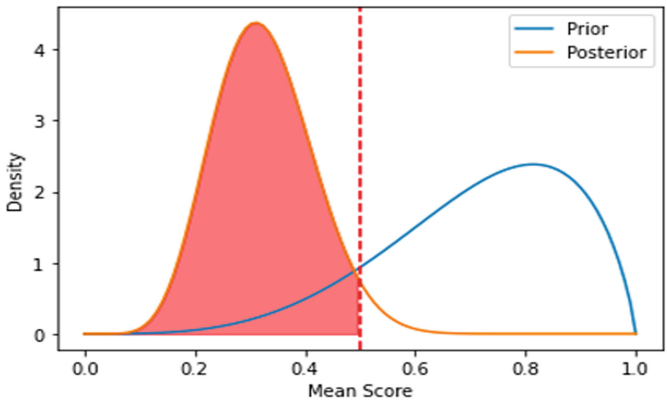

Fig. 4. Posterior distribution for mean score, shown together with the prior distribution from Fig. 3 for comparison. The shaded area represents the probability under the model that the question's mean score is below the decision rule's threshold of 0.5 (dashed vertical line).

In continuous deployment, with regular updating of the posterior distribution, the decision to remove the question could actually have been made in a maximum of 17 student responses (4 correct, 13 incorrect) using the given rule.

5 Conclusion

The use of artificial intelligence to scale effective learning methods is a significant milestone in the advance of educational technology as a whole. However, using automatic question generation at scale presents a new issue of scale: monitoring and taking action on questions. To do this for millions of questions is an impossible manual task in any scenario, so an automated system is a necessary solution. The Content Improvement Service was developed with a systems design perspective to work with the learning platform and the CMS to implement a feedback loop for continuous improvement. The CIS operates at enormous scale—across thousands of textbooks and millions of questions—and yet at the same time it operates at a micro scale—taking action at the individual question level. In the context of Vandewaetere et al.'s [14] adaptive model, the automatically generated questions are both the source and target of adaptation, with the CIS as the pathway to adaptation.

The CIS is a complex system that is a prime example of the technological advances Vandewaetere et al. [14] anticipated arising for the pathway of adaptation. A major objective of the CIS is to make decisions as quickly as possible in order to improve the learning experience for as many students as possible, so complex statistical models are employed to make these determinations efficiently. Yet the statistical methods of the CIS also need to be used in service to a student-centered purpose and in a research-based, grey-box approach. As seen in the examples previously outlined, the CIS uses student feedback as well as question performance data as methods for decision-making. If a question is underperforming, it should not wait to be identified until after potentially hundreds or even thousands of students have experienced it; that question should be

removed and replaced the moment the CIS is confident in that decision. This goal is directly aligned with the student-centered approach of learning engineering. The CIS is the adaptive system that continuously works to optimize the learning resource for students. Furthermore, the CIS itself can adapt and undergo iterative improvement over time, as its methods of analysis are refined and learning science expertise continues to be added.

The CIS presents the opportunity to engage in large-scale data analytics that could reveal new insights in learning science. One clear avenue of future research is to study and improve the AQG process itself. As it is operating on the largest collection of automatically generated questions delivered for student use in natural learning contexts to date, the decisions made by the CIS will provide large, labeled data sets for question quality, which can be used for developing machine learning models to better detect suboptimal questions before they are released. The results of the CIS could also reveal interesting insights into the performance of these automatically generated questions across subject domains. Another avenue of future study is characterizing the timescales needed to optimize questions in the learning environment. Previously, data analysis to identify problematic human-authored questions occurred only after a semester or year of data had been collected, if even at all. In addition to the analysis itself, there is the need for expert review of the results and for any follow-up actions based on that review to be implemented manually. These practical requirements mean students may not receive the benefits of this improvement cycle for a year or longer. With the CIS, these analyses and decisions are improving the learning environment constantly, indicating that at a certain point, every question will have been optimized for students. Discovering this optimization point would be a valuable finding.

References

1. Baker, F.B.: The Basics of Item Response Theory (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation (2001)
2. Durlach, P.J., Ray, J.M.: Designing adaptive instructional environments: Insights from empirical evidence (Technical report 1297). United States Army Research Institute For the Behavioral and Social Sciences (Arlington, VA) (2011). https://apps.dtic.mil/sti/pdfs/ADA552677.pdf
3. Heilman, M., Smith, N.A.: Question generation via overgenerating transformations and ranking (Technical report CMU-LTI-09–013). Carnegie Mellon University, Language Technologies Institute (Pittsburgh, PA) (2009). https://www.lti.cs.cmu.edu/sites/default/files/cmulti09013.pdf
4. IEEE ICICLE: What is Learning Engineering? (2020). https://sagroups.ieee.org/icicle/
5. Kessler, A.: Design SIG Colleagues. Learning Engineering Process Strong Person (2020). https://sagroups.ieee.org/icicle/learning-engineering-process/. Accessed
6. Koedinger, K., McLaughlin, E., Jia, J., Bier, N.: Is the doer effect a causal relationship? How can we tell and why it's important. In: Learning Analytics and Knowledge. Edinburgh, United Kingdom (2016). http://dx.doi.org/https://doi.org/10.1145/2883851.2883957
7. Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. **30**(1), 121–204 (2019). https://doi.org/10.1007/s40593-019-00186-y

8. Martin, F., Chen, Y., Moore, R.L., Westine, C.D.: Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. Educ. Tech. Res. Dev. **68**(4), 1903–1929 (2020). https://doi.org/10.1007/s11423-020-09793-2
9. Sharma, K., Papamitsiou, Z., Giannakos, M.: Building pipelines for educational data using AI and multimodal analytics: a "grey-box" approach. Br. J. Educ. Technol. **50**(6), 3004–3031 (2019). https://doi.org/10.1111/bjet.12854
10. Campenhout, R.: Learning engineering as an ethical framework. In: Sottilare, R.A., Schwarz, J. (eds.) HCII 2021. LNCS, vol. 12792, pp. 105–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-77857-6_7
11. Van Campenhout, R., Dittel, J.S., Jerome, B., Johnson, B.G.: Transforming textbooks into learning by doing environments: an evaluation of textbook-based automatic question generation. In: Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education. CEUR Workshop Proceedings, ISSN 1613–0073, pp. 1–12 (2021). http://ceur-ws.org/Vol-2895/paper06.pdf
12. Van Campenhout, R., Jerome, B., Johnson, B.G.: The impact of adaptive activities in acrobatiq courseware - investigating the efficacy of formative adaptive activities on learning estimates and summative assessment scores. In: Sottilare, R.A., Schwarz, J. (eds.) HCII 2020. LNCS, vol. 12214, pp. 543–554. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50788-6_40
13. Van Campenhout, R. Johnson, B.G., Olsen, J.A.: The doer effect: replicating findings that doing causes learning. In: Proceedings of eLmL 2021: The Thirteenth International Conference on Mobile, Hybrid, and On-line Learning, ISSN 2308–4367, pp. 1–6. https://www.thinkmind.org/index.php?view=article&articleid=elml_2021_1_10_58001
14. Vandewaetere, M., Desmet, P., Clarebout, G.: The contribution of learner characteristics in the development of computer-based adaptive learning environments. Comput. Hum. Behav. **27**(1), 118–130 (2011). https://doi.org/10.1016/j.chb.2010.07.038