AI-Generated Questions in an OER Textbook: Evaluating the Performance of Formative Practice

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

The IAFOR International Conference on Education in Hawaii 2025
Official Conference Proceedings

Abstract
The ubiquitous use of digital learning resources like etextbooks has shifted the learning experience in higher education. Digital learning has led to both new learning tools as well as research in how learning works via the large, rich data sets those digital resources generate. Advances in artificial intelligence have made it possible to develop and scale learning methods—such as formative practice integrated in etextbook content in a learning by doing approach. A primary benefit of generating formative practice is to bring this highly effective learning approach to millions of students using digital textbooks. This paper focuses on an automatic question generation system that has proven to generate effective formative practice for higher education textbooks, as measured through large-scale analyses of question performance metrics and in-classroom implementations. Open education resources (OER), such as OpenStax, offer students and faculty a learning resource without the high cost. In this paper, we evaluate performance metrics such as difficulty and persistence for the automatically generated questions added to an OER textbook for the first time. Used in several large, online chemistry courses at a major public university, this paper showcases the viability of automatically generated questions combined with OER content for increasing the access and affordability of formative practice as a feature in digital textbooks. Key question performance metrics, such as question difficulty and persistence, along with student interaction patterns and behavior, are analyzed, and future applications of OER content with automatically generated questions are discussed.

Keywords: Open Education Resources, Automatic Question Generation, Learning by Doing, Formative Practice

iafor
The International Academic Forum
www.iafor.org

Introduction

Textbooks are a central component of many higher education courses, serving as the primary resource from which instructors expect students to read, learn, and apply knowledge to assignments and assessments. Although textbooks have traditionally been regarded as the gold standard for delivering learning content, they present challenges related to engagement and active learning. First, despite being assigned by instructors, students often do not engage with textbooks as intended (Berry et al., 2010; Burchfield & Sappington, 2000; Connor-Greene, 2000; Schneider, 2001). Data from etextbook platforms have confirmed low reading rates among students, with conventional instructional strategies—such as reading quizzes or discussions—yielding mixed results in increasing engagement (Russell et al., 2023). However, studies indicate that incorporating formative practice is more effective in promoting student engagement than any other reported strategy (Brown et al., 2024). Second, textbooks primarily support passive learning, which is not the most effective approach to knowledge acquisition. Research from Carnegie Mellon University's Open Learning Initiative has demonstrated that embedding formative practice within text content, following a learning-by-doing approach, is about six times more effective than reading alone (Koedinger et al., 2015; Koedinger et al., 2016). This principle, known as the doer effect, has been empirically shown to have a causal relationship with learning outcomes (Koedinger et al., 2016; Koedinger et al., 2018). Further replications of research on the doer effect confirm that this methodology is broadly applicable and should be made available to as many learners as possible (Van Campenhout et al., 2021a; Van Campenhout et al., 2022; Van Campenhout et al., 2023a).

Despite these findings, the integration of formative practice within digital textbook content remains uncommon in higher education. Courseware platforms provide highly effective learning-by-doing experiences but are often challenging to scale due to development costs and adoption barriers. Recent advancements in artificial intelligence have made it feasible to generate the volume of formative practice required for the learning-by-doing model. Automatic question generation (AQG) systems have been increasingly explored by research groups worldwide for a range of educational applications (Kurdi et al., 2020). While various approaches exist for generating and applying these questions, Kurdi et al. (2020) noted that a universally accepted gold standard for automatically generated (AG) questions had yet to be established and more research was needed that used student data to evaluate AG question performance.

An AQG system was developed to generate formative questions directly from textbook content (evaluated in this study). These AG questions were initially integrated into courseware learning environments alongside human-authored questions and evaluated across six courses. The findings indicated no significant differences between AG and human-authored questions in key performance metrics, including engagement, difficulty, persistence, and discrimination (Johnson et al., 2022; Van Campenhout et al., 2021b). Subsequently, these AG questions were embedded into the VitalSource Bookshelf ereader as a study tool called CoachMe. In what is currently the largest known analysis of AG questions using student data, performance metrics from prior evaluations were replicated, confirming these benchmarks at scale (Van Campenhout et al., 2023b). Notably, both the initial comparison of AG and human-authored questions (Van Campenhout et al., 2021b) and subsequent research on CoachMe questions (Van Campenhout et al., 2023b; Van Campenhout et al., 2023c) revealed that the most significant variations in performance metrics were due to the cognitive process dimension of the question type rather than whether the questions were AG or human-

authored. Recognition-based matching questions generally exhibited higher engagement, difficulty indices, and persistence rates, whereas recall-based fill-in-the-blank (FITB) questions tended to show lower averages in these metrics. The distinction between recognition and recall question types has been well-documented for decades (Anderson et al., 2001; Andrew & Bird, 1938), and this research contributes further examples of how these differences influence question performance and student behavior. Additional studies examining student interaction with these questions (Van Campenhout et al., 2023c) and the impact of feedback (Van Campenhout et al., 2024a) have provided new insights into learning behaviors.

Research conducted in natural learning environments is valuable for ensuring external validity and generalizability (Koedinger et al., 2015; Van Campenhout et al., 2023a). Unlike controlled or semi-controlled experiments, classroom-based research does not risk altering natural student behaviors and mitigates ethical concerns related to withholding potentially beneficial learning interventions. While large-scale studies aggregating data from hundreds of thousands of students and millions of answered questions are useful for establishing performance benchmarks, they may not always reflect the nuances of classroom-specific dynamics. In particular, these very large datasets often reflect situations in which the CoachMe questions are optional, whereas instructors who integrate the questions into their courses, e.g., by assigning them directly, tend to see different patterns of student engagement. Studies have shown that students in university courses assigning CoachMe questions engaged with them differently, leading to higher first-attempt accuracy rates and increased persistence (Van Campenhout et al., 2023b; Van Campenhout et al., 2024b), as well as distinct interaction patterns (Van Campenhout et al., 2023c). Since etextbooks are widely used as the primary learning resource in many courses, it is essential to examine how AG questions function in real classroom settings. Given the substantial influence that course context and instructor implementation strategies can have on student engagement and learning, understanding their impact on AG question usage is a critical area for further investigation (Kessler et al., 2019; Van Campenhout & Kimball, 2021).

Open Educational Resources (OER) have emerged in recent decades as an alternative learning resource option in higher education, providing freely accessible and openly licensed learning materials aimed to help eliminate financial barriers for students. The rising costs of textbooks have been shown to negatively impact student access to required course materials, with many students opting to forgo purchasing textbooks due to affordability concerns (Raneri & Young, 2016; Nagle & Vitez, 2019). While OER offers one path to mitigate this issue, OER has not received widespread support from faculty and administrators, with content quality being a major concern cited (Raneri & Young, 2016; OnCampus Research, 2024). Studies investigating its use in higher education have demonstrated that courses using OER can yield comparable (or in some cases even improved) learning outcomes when compared to commercial textbooks, that faculty perceived equal preparedness from students, and that overall student and faculty perceptions were positive (Bliss et al., 2013; Clinton & Khan, 2019； Fischer et al., 2015; Hilton, 2016). Hilton's synthesis of research published between 2015 and 2018 further supports these findings, showing that OER adoption does not negatively impact student learning and is generally perceived positively by both students and faculty (2020). A meta-analysis by Clinton and Khan (2019) also found OER adoption was associated with lower course withdrawal rates, indicating potential benefits for student retention. Similarly, research by Colvard, Watson, and Park (2018) found that students using OER had higher course grades and lower DFW (drop, fail, withdraw) rates, with the greatest benefits observed among historically underserved student populations. While OER adoption

at universities has been slowly increasing, there are still barriers to OER adoption, such as instructor concerns about quality and lack of insight into OER availability and adoption paths (OnCampus Research, 2024).

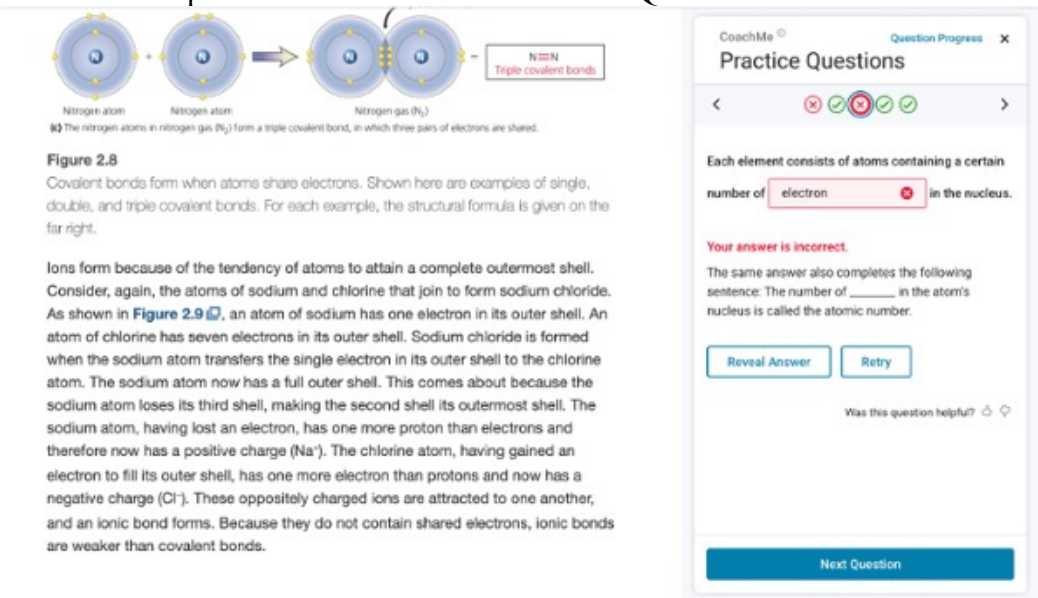
OER textbooks are passive learning environments—the same as traditional textbooks—and can often have fewer opportunities for interactive learning compared to commercial textbooks. The integration of OER with additional technologies—such as automatic question generation—offers new opportunities to enhance student engagement and support personalized learning experiences. This study aims to extend the existing research on AG question performance in classroom settings by examining questions generated from an OpenStax chemistry textbook. Faculty at a major public university used the OER textbook with the AG questions in a fall semester of Chemistry 101 and a spring semester of Chemistry 101 and Chemistry 102. With a combined 1,555 students and 30,529 questions answered, the data collected from these courses provide insight into both implementation practices and student behaviors as well as question performance metrics for OER textbooks, providing necessary insights into the quality of questions generated from this source content.

Methods

Formative Practice

The AQG system utilized in this study is a rule-based system designed by experts. Neither of the question types examined in this paper (matching and FITB) was generated using large language models. Instead, the system processes the course textbook as its corpus for natural language processing and by leveraging both syntactic and semantic information, it identifies key sentences and important terms, which are then transformed into questions through a structured set of rules (for further details on the AQG system, see Van Campenhout et al., 2021b; Van Campenhout et al., 2023b). Once generated, these questions are embedded alongside the corresponding textbook section. As illustrated in Figure 1, students receive immediate feedback upon submitting an answer. For FITB questions, scaffolding feedback incorporating an additional example from the textbook has been shown to be the most effective in promoting student persistence and improving second-attempt accuracy rates (Van Campenhout et al., 2024). When students respond incorrectly, they have the option to either retry the question—resetting it in the process—or reveal the correct answer, with additional retry attempts available. A progress panel allows students to track their completion percentage, view the correctness status of each question, and navigate between different questions and question sets. Prior research has demonstrated that this progress panel enhances student motivation, encouraging them to complete more, if not all, of the available required practice (Van Campenhout et al., 2023d).

Figure 1: An Example of an FITB Formative Practice Question With Immediate Feedback



As students interact with the etextbook and answer questions, the ereader platform continuously collects clickstream data, assigning timestamps to each user action. This fine-grained, contextual microlevel data is highly valuable for educational data science (Fischer et al., 2020; Van Campenhout & Johnson, 2023e), enabling researchers to address both longstanding and emerging questions in education (McFarland et al., 2021). In this study, we leverage this data to investigate both well-established issues—such as textbook engagement and the benefits of formative practice—as well as the evolving research area of automatic question generation.

Implementation

The Chem 101 and 102 courses were both run at a major public university. In fall 2023 (F23) Chem 101 was co-taught by two faculty, as the course was a combination of many sub-sections and consisted of more than 700 students. Deployed as an online synchronous course, students were expected to attend lectures and do several different types of assignments each week. Reading chapters from the textbook was an expectation; however, not all sections of each chapter were included in the assigned reading and the practice was incentivized but not assigned. If students completed a total of 65% of the practice in the assigned chapters (1–9) by the end of the course, they could drop their lowest reading quiz score. In spring 2024 (S24), the same faculty members no longer co-taught one class, but rather taught either the repeat section of Chem 101 or the continuation of Chem 102. Chem 101 included two more chapters (1–11) than the F23 section. Chem 102 continued from the F23 Chem 101 course with chapters 10–17 assigned.

While these three course sections are from the same university and involve the same instructors, it is not realistic to compare the fall and spring semesters directly for several reasons. First, the change from co-teaching to teaching separately still introduces differences in course policy and delivery of content that affects student behavior. Second, fall and spring semesters are often comprised of different student cohorts with different characteristics (as will be seen in the results analysis), which also restricts direct comparison. Lastly, the question performance cannot be directly compared even for the Chem 101 sections as the questions were changed between semesters. An advantage of the AQG system is that it can

be updated over time with iterative improvements and textbooks can be "rerun" to produce a new question set. Between these semesters, a book rerun was scheduled that updated the question set prior to student use in the spring. However, these comparisons while knowing these changes does shed light on some interesting student behaviors.

Results

As has been seen in other implementations of this formative practice, not assigning the practice leads to depressed overall engagement (Van Campenhout et al., 2024b). F23 Chem 101 averaged 15.8 questions answered per student. S24 Chem 101 had 37.6 questions per student, more than double the F23 course. Chem 102 was similar to F23 Chem 101 with 16.3 questions per student. However, this average per student is not representative of student engagement behavior, as it is more common for some students to do all the practice and some do none. In F23 Chem 101, 47% of students did questions and 53% did none; in S24 Chem 101, 62% of students did questions and 38% did none; in S24 Chem 102, 53% of students did questions and 47% did none.

Difficulty and persistence are two performance metrics that provide insight into the question performance. Difficulty in this instance refers to the difficulty index, where a higher value means more students answered the questions correctly and a lower value means fewer students answered the questions correctly. Table 1 shows the mean values for matching questions and FITB questions for each section. Consistent with prior research (Van Campenhout et al., 2023), the matching questions have a higher difficulty mean than the FITB questions (meaning students get the matching correct on their first attempt more frequently than the FITB). What is unusual is the differences in difficulty between course sections. The F23 Chem 101 section has very low difficulty means. This is surprising as research on classroom implementations have found difficulty means to be much higher than the aggregated dataset means (Van Campenhout et al., 2023b; Van Campenhout et al., 2024b). It is difficult to discern the exact reason for this given the complexity of the context—it could be related to the course assignment policy, the strategy and motivation of the students who chose to answer, or the questions. As an internal validation step, the research team did review the question set to ensure this question set was not an outlier compared to other textbooks.

Even more interesting is the dramatic difference in difficulty means for the S23 Chem 101 course. With a mean of 83.88% for matching and 79.31% for FITB, this course is consistent with prior research on classroom implementations. The textbook was the same, but it was a different semester, different student cohort (possibly with different motivational characteristics), one instructor instead of two, and a different question set. Due to these variations, it is not possible to ascertain the specific reason for the difference, but it does eliminate concerns that an OER title might not be suitable for quality AG questions. The S24 Chem 102 course had difficulty means between the F23 and S24 Chem 101 means.

When students answer questions incorrectly on their first attempt, persistence is the rate that students continue to answer a question until they reach the correct response. Persistence is therefore a subset of the difficulty data set. Prior research has shown that persistence is higher for matching (recognition type) than FITB (recall type) (Van Campenhout et al., 2023b; 2024b), which is consistent with these results, shown in Table 1. The same course trends seen for difficulty are seen for persistence; F23 Chem 101 has the lowest persistence rates while S24 Chem 101 had the highest persistence rates.

Table 1: Difficulty and Persistence Metrics

Section	Students	Question Total	Total Answered	Matching Mean	Matching Persistence	FITB Mean	FITB Persistence
F23 101	744	119	11,769	61.94	52.54	43.27	48.64
S24 101	260	403	9,774	83.88	83.82	79.31	77.31
S24 102	551	207	8,986	73.39	63.01	59.69	62.14

Formative practice is intended to support the learning process and therefore assigning it is based on completion and not first attempt accuracy, as that would negate the goal of formative practice as low or no-stakes practice. However, instructor concerns about student behaviors around cheating are valid. How do we know students are taking the practice seriously and not just inputting garbage at the last minute to get their points? To investigate this, we identified a set of rules to analyze the FITB responses that capture the majority of responses deemed "non-genuine," meaning not a legitimate attempt at the correct response. This includes responses under three characters, punctuation, no vowels, and known responses such as "idk." The non-genuine response rate for the aggregated big data set was 12% (Van Campenhout et al., 2023), but this rate varies in classroom contexts (Van Campenhout et al., 2024b).

Table 2 shows the non-genuine response rates were highest for F23 Chem 101 at 15.4% and lowest for S24 Chem 101 at 5%, with S24 Chem 102 in the middle at 10%. These percentages are consistent with the trends of difficulty and persistence for these courses, and are also clustered around the aggregated non-genuine response rate. Once the non-genuine response rate is calculated, this set of questions is further analyzed for persistence. For the questions that students input a non-genuine response, how often do those students persist in submitting the correct answer? In all courses, it is higher than the aggregated rate of 46%, with F23 Chem 101 coming in the lowest at 53.17% of the time. S24 Chem 102 was in the middle at 68.33% and S24 Chem 101 was the highest at 85.05%. These persistence rates indicate that the majority of students who input a non-genuine response may have done so as a strategy in order to see feedback or request an answer reveal. These students went through the effort of retrying and entering the correct response which indicates they were not merely trying to game the system just for points.

Table 2: Non-genuine Response Rates and Persistence

Section	Students	Question Total	Total Answered	FITB Non-Genuine Answers	FITB Non-Genuine Persistence
F23 101	744	119	11,769	15.4%	53.17%
S24 101	260	403	9,774	5.0%	85.05%
S24 102	551	207	8,986	10.0%	68.33%

These courses provide another unique opportunity for analysis, as there are students who took F23 Chem 101 who either moved on to S24 Chem 102 or retook S24 Chem 101. There were 427 of 744 students who continued to S24 Chem 102 and 81 students who retook Chem 101 in S24. In Tables 3 and 4, we investigate the difference in matching and FITB question difficulty means by students who would continue to Chem 102 or retake Chem 101.

In F23 Chem 101, the 427 students who would persist to S24 Chem 102 had slightly higher matching mean difficulty than their peers, but a slightly lower FITB mean difficulty.

Interestingly, once in the Chem 102 course, the students who had continued from Chem 101 had a *lower* mean first attempt for both matching and FITB than their peers. It is unclear why this is, but it is reasonable to connect the low mean difficulty from F23 Chem 101 and the lower mean scores with the students who continued on as a related outcome, perhaps as a characteristic of the cohort or level of importance they placed on the optional assignment.

The data for students who retook Chem 101 tell a different story. In F23 Chem 101, students who would retake the course had substantially lower mean first attempt difficulty on both questions than their peers. This is not surprising given their retake status. However, those same students in S24 Chem 101 then outperform their peers on both question types. The difference for the FITB questions is dramatic: 28.74% in the fall compared to 86.72% in the spring. Those students clearly decided to make better use of the questions with more effort on their first attempts.

Table 3: Question Difficulty for Students Who Continued From Chem 101 to Chem 102

Question Type	Student Persisted	F23 Total Answered	F23 Mean	S24 Total Answered	S24 Mean
Matching	Persist	3960	62.95	2616	72.17
	N/A	1842	59.77	607	78.58
FITB	Persist	4088	42.37	4695	58.64
	N/A	1879	45.24	1068	64.33

Table 4: Question Difficulty for Students Who Retook Chem 101

Question Type	Student Persisted	F23 Total Answered	F23 Mean	S24 Total Answered	S24 Mean
Matching	Retake	528	50.00	297	85.16
	N/A	5274	63.14	18.12	83.66
FITB	Retake	581	28.74	1009	86.72
	N/A	5386	44.84	6656	78.19

One final metric useful for gaining insight into student perceptions of the questions is the thumbing rate. After answering a question, students have the option of providing a thumbs up or thumbs down on the question. This thumbing data is used by a platform-wide adaptive content improvement system that uses the thumbing data to determine if questions should be removed and replaced (Jerome et al., 2022). Using a dataset of over 3,594,408 answered questions, the overall thumbs down rate was 1.9% per thousand questions and the thumbs up rate was 3.3% per thousand questions. In these semesters (Table 5), the thumb up rates were much higher—between 48 and 57 per thousand. The thumbs down rate was between 1 and 6 per thousand. Instructors did point out the thumbing option but there were no expectations for student thumbing, so the high thumbs up rates indicate overall student satisfaction. It's also noteworthy that the highest thumbs up rate was in F23 Chem 101, which had the lowest mean difficulty on questions.

Table 5: Thumbing Rates per 1,000 Sessions by Course

Section	Thumbs Up	Thumbs Down
F23 101	56.59	5.86
S24 101	51.38	1.74
S24 102	48.64	2.56

Conclusion

This study provides strong evidence supporting the effectiveness of automatically generated questions as a scalable and viable solution for integrating formative practice into Open Educational Resource (OER) textbooks. Across multiple large-scale chemistry courses, AG questions performed comparably to prior research in key performance metrics such as difficulty and persistence. The positive student response as reflected in the high thumbs-up ratings for AG questions further validates their perceived value in the learning process. While the course policy of not assigning the questions depressed overall engagement, the spring 2024 Chemistry 101 course demonstrated significantly higher student engagement, with students answering more than twice the number of questions compared to the fall 2023 section, suggesting that course design and instructional context play a crucial role in shaping student behavior. The ability to analyze students who moved from Chem 101 to Chem 102 or retook Chem 101 gave a unique view into student behaviors and performance. Notably, students who needed to retake Chem 101 performed dramatically better on the formative practice, indicating a new motivation for taking advantage of learning tools.

The successful addition of AG questions within an OER textbook underscores the potential for AI-driven formative practice to enhance the student learning experience in addition to access and affordability in higher education. These findings align with broader research on OER efficacy, which has consistently shown comparable or improved student outcomes compared to traditional textbooks while reducing financial barriers.

As digital learning environments continue to evolve, integrating AI-generated formative practice within any content, including OER, presents new opportunities for improving student engagement and learning outcomes. Future research should explore how different instructional strategies impact student interaction with AG questions and how that supports diverse learner needs. The findings from this study contributes evidence demonstrating that AG questions are an effective solution for expanding high-quality formative practice within OER textbooks, as previously demonstrated with commercial publisher textbooks.

References

Anderson, L. W., Krathwohl, D. R., 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).* Longman.

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

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

Bliss, T. J., Robinson, T. J., Hilton, J., & Wiley, D. (2013). An OER COUP: College teacher and student perceptions of Open Educational Resources. *Journal of Interactive Media in Education, 2013*(1), 4.

Brown, N., Van Campenhout, R., Clark, M., & Johnson, B. G. (2024). Are students reading? How formative practice impacts student reading behaviors in etextbooks. *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S'24)*, 383–387. https://doi.org/10.1145/3657604.3664668

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

Clinton, V., & Khan, S. (2019). Efficacy of open textbook adoption on learning performance and course withdrawal rates: A meta-analysis. *AERA Open, 5*(3). https://doi.org/10.1177/2332858419872212

Colvard, N. B., Watson, C. E., & Park, H. (2018). The impact of open educational resources on various student success metrics. *International Journal of Teaching and Learning in Higher Education, 30*(2), 262–276. https://eric.ed.gov/?id=EJ1184998

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

De los Arcos, B., Farrow, R., Perryman, L.-A., Pitt, R., & Weller, M. (2014). *OER research hub: Evidence report 2013-2014.* OER Research Hub.

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education, 44*(1), 130-160. https://doi.org/10.3102/0091732X20903304

Fischer, L., Hilton, J., Robinson, T. J., & Wiley, D. (2015). A multi-institutional study of the impact of open textbook adoption on the learning outcomes of post-secondary students. *Journal of Computing in Higher Education, 27*(3), 159-172. https://doi.org/10.1007/s12528-015-9101-x

Hilton, J. (2016). Open educational resources and college textbook choices: A review of research on efficacy and perceptions. *Educational Technology Research and Development, 64*(4), 573-590. https://doi.org/10.1007/s11423-016-9434-9

Hilton, J. (2020). Open educational resources, student efficacy, and user perceptions: A synthesis of research published between 2015 and 2018. *Educational Technology Research and Development, 68*(3), 853–876. https://doi.org/10.1007/s11423-019-09700-4

Jerome, B., Van Campenhout, R., Dittel, J. S., Benton, R., Greenberg, S., & Johnson, B. G. (2022). The Content Improvement Service: An adaptive system for continuous improvement at scale. In Meiselwitz, et al., Interaction in New Media, Learning and Games. HCII 2022. Lecture Notes in Computer Science, vol 13517, 286–296. Springer, Cham. https://doi.org/10.1007/978-3-031-22131-6_22

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

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

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. *Proceedings of the Second ACM Conference on Learning@Scale.* http://dx.doi.org/10.1145/2724660.2724681

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

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.*http://dx.doi.org/10.1145/2883851.2883957

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*(1), 121-204. https://doi.org/10.1007/s40593-019-00186-y

McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education data science: Past, present, future. *AERA Open, 7*(1), 1-12. https://doi.org/10.1177/23328584211052055

Nagle, C., & Vitez, K. (2019). *Fixing the broken textbook market: Third edition.* U.S. PIRG Education Fund.

OnCampus Research. (2024). *Faculty Watch: Attitudes and behaviors toward course materials.* https://www.oncampusresearch.org/faculty-watch

Raneri, A., & Young, L. (2016). Leading the Maricopa Millions OER Project. *Community College Journal of Research and Practice, 40*(7), 580–588. https://doi.org/10.1080/10668926.2016.1143413

Russell, J.-E., Smith, A. M., George, S., & Damman, B. (2023). Instructional strategies and student etextbook reading. *ACM International Conference Proceeding Series*, 613–618. https://doi.org/10.1145/3576050.3576086

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

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

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

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

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

Van Campenhout, R., Jerome, B., & Johnson, B. G. (2023a). The doer effect at scale: Investigating correlation and causation across seven courses. *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023).*https://doi.org/10.1145/3576050.3576103

Van Campenhout, R., Jerome, B., & Johnson, B. G. (2023e). Engaging in student-centered educational data science through learning engineering. In A. Peña-Ayala (Ed.), *Educational data science: Essentials, approaches, and tendencies*(pp. 1–40). Springer. https://doi.org/10.1007/978-981-99-0026-8_1

Van Campenhout, R., Jerome, B., Kimball, M., Clark, M., Dittel, J. S., & Johnson, B. G. (2024). An investigation of automatically generated feedback on student behavior and learning. *LAK '24: Proceedings of the 14th Learning Analytics and Knowledge Conference*, 850–856. https://doi.org/10.1145/3636555.3636901

Van Campenhout, R., Johnson, B. G., & Olsen, J. A. (2021a). The doer effect: Replicating findings that doing causes learning. *Proceedings of eLmL 2021: The Thirteenth International Conference on Mobile, Hybrid, and Online Learning.*https://www.thinkmind.org/index.php?view=article&articleid=elml_2021_1_10_58001

Van Campenhout, R., Johnson, B. G., & Olsen, J. A. (2022). The doer effect: Replication and comparison of correlational and causal analyses of learning. *International Journal on Advances in Systems and Measurements, 15*(1&2), 48–59. https://www.iariajournals.org/systems_and_measurements/sysmea_v15_n12_2022_paged.pdf

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 – Hawaii 2021 Official Conference Proceedings*, 151–161. https://doi.org/10.22492/issn.2189-1036.2021.11

Van Campenhout, R., Selinger, M., & Jerome, B. (2023b). Designing a student progress panel for formative practice: A learning engineering process. *Proceedings of the Third Annual Meeting of the International Society of the Learning Sciences.* https://2023.isls.org/proceedings/

Contact email: rachel.vancampenhout@vitalsource.com