AI-Generated Formative Practice and Feedback: Performance Benchmarks and Applications in Higher Education

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

Abstract

Integrating formative practice questions with text content is a highly effective learning method. Millions of AI-generated formative practice questions, embedded in thousands of publisher e-textbooks, are now available to students in higher education. This paper reviews findings from a multi-year research program to synthesize performance benchmarks for automatically generated questions and feedback derived from large-scale student interaction data. In addition, we report classroom-based applications that demonstrate how these questions can support learning when integrated into instruction. A central contribution of this review is to identify barriers to effectively scaling student engagement with formative practice, identifying both the successes of automatic question generation systems and the persistent challenges that must be addressed to maximize their potential for classroom impact.

1 Introduction

Formative practice has long been known to be highly effective for learning for students of all ages, but especially struggling students [1, 2]. Research studying the relationship between integrating formative practice with expository content and learning outcomes in digital learning environments found that doing practice was an average of six times more effective for learning than just reading [3, 4]. Called the doer effect, this learning science principle was also shown to have a causal impact on learning [4, 5]. Studies replicating the doer effect in different learning environments confirmed generalizability of this learning by doing approach [6, 7]; however, bringing this method to more students was a persistent challenge. Artificial intelligence presented a solution to this challenge as tools became robust enough to develop an automatic question generation (AQG) pipeline capable of generating millions of practice questions in very little time. The primary objective of the AQG system was to generate formative practice and feedback to be placed alongside textbook content in an ereader platform for use by students in higher education contexts. After the release of the automatically generated (AG) questions, years of research looking at millions of student-question interactions contributed to setting performance metric benchmarks for AG questions and revealed new insights into student behaviors and learning [8-12].

The aim of this paper is twofold. First, we synthesize findings from our multi-year program of research on AI-generated formative practice questions, highlighting both the technical performance benchmarks and their impact in classroom contexts. Second, we reflect on the persistent challenges of effectively scaling student engagement with formative practice, setting out a forward-looking vision for integrating these tools into everyday learning. By combining a review of empirical results with an analysis of practical barriers, we seek to show not only that AI-generated practice can achieve comparable quality to human-authored questions, but also how these systems can maximize learning potential when thoughtfully embedded into teaching and learning environments.

In line with this dual focus, the paper is organized to address both performance at scale and applications in authentic classrooms. Performance metrics drawn from millions of student-question interactions establish validity and reliability of AG questions, while classroom-based studies demonstrate how instructor course policies and student use patterns influence outcomes. Together,

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 337–344

these complementary perspectives underscore how AQG contributes to learning when implemented in real-world educational settings and highlight the remaining obstacles to broader adoption.

2 AQG Methods

The AQG system is designed to support students with formative practice while engaging with textbook material, and so to ensure the questions are closely aligned to the source content, the AQG system uses the textbook as the corpus for natural language processing. Kurdi et al. [13] recommended describing the system according to level of understanding and procedure of transformation. In this system, the level of understanding includes both syntactic and semantic information, and the procedure of transformation is primarily rule-based.

Natural language processing tasks are executed using the spacy library [14], employing its CPU-optimized large language model (en_core_web_lg). Question generation relies on both syntactic and semantic understanding of the text. For cloze question types, this dual-level analysis enables two central operations: identifying the sentences from which questions will be generated and selecting the term(s) to be removed as answers. Syntactic information, including part-of-speech (POS) tagging and dependency structure, informs both content sentence selection and answer word identification. Additionally, semantic information contributes to recognizing conceptually important material. The transformation process that converts sentences into questions follows a rule-based approach developed by experts.

To identify high-value sentences, the textbook is segmented into logical sections of roughly 1,500 words, following the major organizational structure of the textbook such as chapters and their primary headings; sections exceeding this length are further subdivided. Within each section, sentence importance is assessed using the TextRank algorithm [15]. TextRank evaluates similarity between sentences by computing their vector embeddings, the effectiveness of which depends on the embedding technique employed. Our implementation uses a word2vec-based model [16] within spacy, which forms sentence embeddings by averaging the token vectors in each sentence. Prior to embedding, the AQG system filters out stop words and non-alphabetic tokens (e.g., punctuation, numerals). Sentences that are overly short (under 5 words) or long (over 40 words) are also excluded, as they are generally less appropriate for question formation. TextRank is applied to the remaining sentences in each section.

A second core aspect of cloze question creation involves selecting the appropriate answer word(s) from the previously identified sentences. Our system accounts for multiple factors in this process, such as word frequency within the corpus and whether a term appears in the textbook's glossary. However, the most critical factor is part of speech: only nouns and adjectives are considered viable candidates for answer blanks. Research from authentic learning contexts supports this focus—questions that target these parts of speech tend to receive better evaluations from learners than those using verbs or other word types [17]. As such, POS tagging is a fundamental component of AQG, as it is in many NLP applications.

Multiple choice or glossary term compare-and-contrast questions rely on the existence of a textbook glossary, but are created using similar methods.

This AQG approach is designed for broad applicability across academic disciplines but is not suitable for all subject areas; notably, it is not effective for mathematics or language instruction.

Feedback is provided using textbook sentences that are related to the one from which the question stem was created—either a different sentence containing the same answer term (example in Figure 1) or neighboring sentences that provide added context. Outcome-based feedback (correct/incorrect) is always presented.

While the system does not attempt to calibrate question difficulty during generation, student response data collected after deployment is used to monitor difficulty levels. Questions identified as excessively difficult for formative purposes are automatically replaced [17]. Paraphrasing or rewording of textbook content is intentionally avoided to ensure terminology consistency between questions and the source material. The resulting questions with integrated feedback are delivered in clusters that open alongside the relevant textbook section and allow students to get immediate feedback, retry or reveal answers, and rate questions (Figure 1).

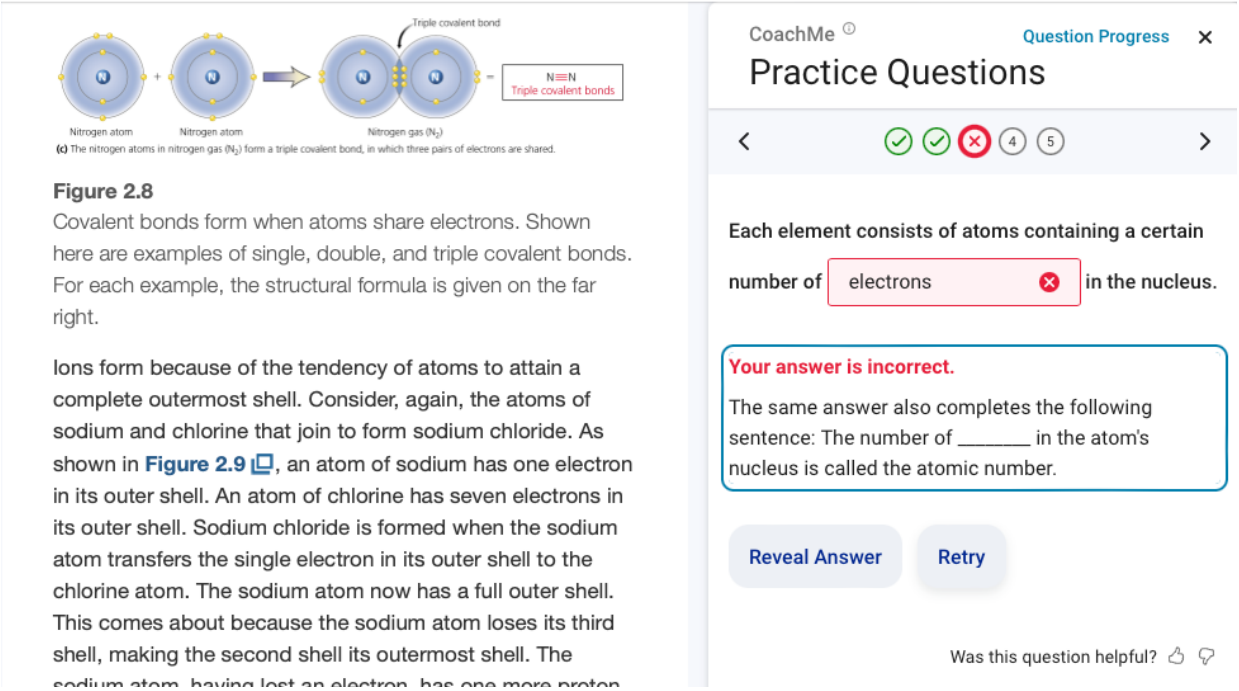

Figure 1. A fill-in-the-blank question open next to the textbook content

The original AQG system was developed without the use of large language models (LLMs) for two key reasons. First, LLMs lacked sufficient reliability at the time of the pipeline's development. Second, their potential to introduce factual inaccuracies posed an ethical concern, especially given the vast number of questions being generated—making human oversight unfeasible at scale. However, LLMs have key strengths that could potentially be harnessed for specific tasks within the existing AQG pipeline [18] or providing error-specific feedback on open-ended questions [19]. While crafting open-ended questions is relatively straightforward, offering targeted feedback is significantly more complex. Intelligent tutoring systems are known for delivering highly effective, individualized feedback that addresses student errors, making them among the most impactful forms of computer-based learning [20, 21]. Historically, scaling this type of feedback has been a major limitation. However, the proficiency of LLMs in text comparison may offer a viable path forward in addressing this challenge.

In the autumn of 2024, two new types of open-ended questions were introduced alongside the existing AG formative question types: a glossary term compare-and-contrast prompt and a "write your own exam question" task. These additions were chosen specifically to engage learners in advanced cognitive process dimensions [22]. To support these questions, an LLM is employed to analyze student responses by comparing them to the corresponding textbook sections and generating constructive, personalized feedback. Although the rule-based AQG pipeline had the capacity to formulate such open-ended prompts previously, deploying them without the ability to provide feedback risked leaving students unsure about the accuracy of their answers—potentially reinforcing misconceptions. As a result, implementing these question types necessitated the inclusion of a mechanism for tailored feedback.

3 Performance Metric Benchmarks

A benefit of digital learning environments is their ability to collect enormous quantities of high-quality data [23]. These microlevel clickstream data allow us to investigate old questions with novel data and gain a finer-grained understanding of student learning processes [24, 25]. The microlevel data collected by the ereader platform are valuable for investigating both the performance of AG questions and student behaviors. The platform records each student interaction with a timestamp and unique numeric identifier for the student. Student-question sessions are formed by grouping all interactions of a single student on a single question. No personally identifiable information is collected by the platform. These data

are then used to evaluate several different performance metrics, including:

- Difficulty index: Percentage of sessions in which the student's first answer attempt was correct (lower values correspond to more difficult questions).
- Persistence rate: Among sessions in which the first attempt was incorrect, the percentage in which the student continued until submitting a correct answer.
- Thumbs up rate: Number of thumbs up ratings per 1,000 student-question sessions (one rating opportunity per session).
- Thumbs down rate: Number of thumbs down ratings per 1,000 student-question sessions.

The initial release of AG questions for student use was in a courseware environment where AG questions were intermixed with human-authored questions and placed intermittently with short content lessons. This first research found no difference in how students use AI-generated versus human-authored questions. Comparing automatically generated questions to human-authored questions in the same course using a mixed-effects logistic regression model found they were similar on engagement, difficulty, persistence, and discrimination [8, 9]. The most notable difference was in the cognitive process dimension of the questions: recall types and recognition types grouped together on performance metrics—regardless of method of creation.

With satisfactory performance in a courseware environment, the AG questions were then delivered as a free study feature (CoachMe) in the Bookshelf ereader, deploying millions of questions across thousands of textbooks. Analysis of over 7 million student-question interactions confirms these performance metric benchmarks at scale—recognition-type questions are generally easier than recall-type questions and have higher persistence. Investigating student answers revealed insight into behaviors: only about 12% of students-question interactions had a "non-genuine" input to the fill-in-the-blank, and nearly half of those students persist in answering until they get the correct response, indicating non-genuine responses were part of a strategy for many students [26]. Tracing interaction patterns also revealed the type of question impacted how students engaged with them [27]. The scale of this release made human monitoring of question performance impossible, so a content improvement service (CIS) was developed. The CIS is a platform-level adaptive system that monitors every student-question interaction in real time and deploys tools such as Bayesian evaluation of difficulty metrics and student ratings (thumbs down specifically) to determine if questions need to be removed and replaced [28]. Across a total of 3,594,408 question sessions, the overall thumbs down rate observed was 1.94 [29].

To provide an updated set of aggregated performance metrics, all student-question interaction events were retrieved starting from the feature's launch on January 1, 2022, to June 11, 2025. The resulting dataset consisted of 16,645,791 sessions across 2,485,201 unique questions, 822,678 students, and 14,371 textbooks, with a total of 26,169,711 interaction events. Table 1 summarizes these performance metrics by question type.

Compared to the performance metrics from [26] in 2023, the overall trends by the cognitive process

	Answered	Mean Difficulty	Persistence	Thumbs Up Rate	Thumbs Down Rate
Matching	4,028,835	80.3	72.8	3.56	1.00
Self-Graded Submit & Compare	526,080	86.8	NA	4.64	2.37
FITB	11,912,905	61.4	62.1	3.28	1.73
Multiple Choice	205,774	74.1	76.1	3.68	2.10

Table 1. Performance metrics by question type.

dimension of the question types remain the same. The recognition-type matching and multiple choice questions are easier and have higher persistence than the recall FITB type. However, we see some interesting changes. In 2023, the FITB had a difficulty of 54.7 and persistence of 58.5. The most recent data show an increase for both metrics to a difficulty of 61.4 and persistence of 62.1. This increase is overall positive, and potentially was impacted by improvements made to the AQG pipeline and question placement in December of 2023. The only other large difference is persistence for multiple choice, which fell from 93.6 to 76.1—potentially related to the nearly tenfold increase in data collected on this question type since 2023.

In addition to monitoring performance benchmarks of the AG questions themselves, we investigated AG feedback. The type of feedback used for formative practice matters. Scaffolding feedback that provides another context (Figure 2) was most effective for increasing student persistence in answering until correct as well as decreasing the time it took to get to the correct answer [11]. Additionally, the advances in large language models (LLMs) made it possible to scale personalized, error-specific feedback for open-ended question types—a hallmark feature of intelligent tutoring systems [19]. Introducing LLM-based error-specific feedback for open-ended questions produced by this AQG pipeline provided experience with an LLM-based feature that could replicate the hallmark personalized feedback of intelligent tutoring systems but required careful development to minimize potential LLM failures [17].

4 Data from the Classroom

The millions of questions available for analysis provide valuable performance benchmark metrics for AG questions. However, the large aggregated dataset includes all learners in all learning contexts—even those who only answered a few questions. Therefore, it was also valuable to engage in classroom-based research to determine how instructor course policies impacted student engagement with the practice and how the AG formative practice might impact learning. Studying 19 course sections where faculty assigned these AG questions as a participation homework assignment showcased how classroom contexts and course policies increased student engagement and impacted performance metrics [29]. Nearly all students answered 100% of the questions, even when only 80% was required to receive credit. Across all courses, the matching questions had a mean difficulty of 82.8% and a persistence of 96.7%. The FITB questions had a mean difficulty of 82.7% and a persistence of 94.0%. The higher difficulty index and persistence for questions in the classroom setting indicates students put more effort into their first attempt at the question and were motivated to continue answering until they input

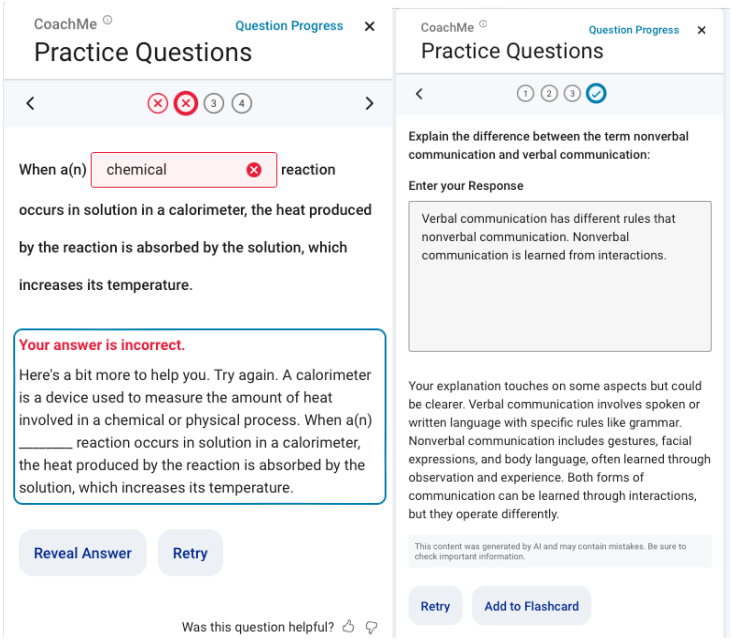

Figure 2. Scaffolding feedback for FITB questions (left) and LLM-based error-specific feedback for open-ended questions (right).

the correct response. The non-genuine responses for FITB ranged widely between courses, but remarkably, 12 of 19 courses had persistence over 99% for non-genuine responses. Faculty observed increased preparedness for classroom discussions and higher quality written assignments and projects and students anonymously reported finding the practice helpful for both learning and accountability on course evaluations [12].

In two semesters of a large cognitive psychology course, a change in faculty policy similarly shifted students from doing practice at the end of the course when it would not be helpful for the exams to prior to the related exam [30]. This change led to a statistically significant increase in exam scores (particularly meaningful for struggling students at the 25th and 50th percentile). Additionally, a post hoc analysis replicating Koedinger et al.'s doer effect analysis found results consistent with the literature. This first analysis of AI-generated questions eliciting the same doer effect principle in the classroom confirms the utility of AI for question generation at scale [30].

5 Recommendations, Challenges, Future Work

A key contribution of this review is to identify not only what the AQG pipeline has achieved in terms of question quality and learning outcomes, but also the persistent barriers that hinder scaling student engagement with formative practice. Each individual research study conducted on this AQG system since its initial release in 2019 investigates specific components in detail, such as performance metrics, student perceptions, feedback, student engagement patterns, textbook reading, learning outcomes, etc. Together, this rigorous evaluation of nearly every aspect of question performance and student behaviors and learning is essential to a comprehensive overview of the efficacy of AI-generated questions for formative practice at scale.

While our analyses confirm that AG questions perform well across multiple metrics and can replicate the doer effect in classroom settings, two persistent barriers emerge. First, faculty awareness and adoption remain uneven—many instructors are not fully informed about the availability of AG questions embedded in their textbooks. Second, student engagement is highly dependent on course structures; voluntary use of AG practice is typically low unless supported by meaningful course incentives or policies. These barriers illustrate that successful application of AQG in classrooms is not a purely technical challenge but an educational and organizational one. Addressing these barriers is essential to realizing the potential of formative practice: maximizing learning through classroom application. Without meaningful faculty engagement, voluntary student use of the questions will remain low. Instructors remain the most meaningful agents of change in the classroom and helping to inform and educate instructors as key partners in implementation will remain the focus of future efforts.

Future work will always include iterative improvement to the AQG pipeline. The analysis of the questions showcases their validity, yet continued refinement can further improve question quality. We have evidence of the importance of this improvement cycle, as changes made to sentence selection and placement within the text in the winter of 2023 resulted in a reduction of thumbs down ratings from 1.95 to 1.39 per thousand. While the thumbs down rate is very low, decreasing it by more than 25% indicates an effective improvement that could influence student perceptions of the questions. While LLMs were not used in the existing AQG pipeline, we have conducted promising research on how introducing LLMs at key steps in the pipeline could further increase question quality [18].

Taken together, the results of this research establish clear performance benchmarks for AI-generated formative practice questions, demonstrating that they perform comparably to human-authored questions across difficulty, persistence, and engagement metrics at scale. Classroom-based implementations further confirm that when these questions are embedded into instruction, they not only support higher persistence and accuracy but also contribute to measurable gains in exam performance and student preparedness. These findings underscore that AI-generated formative practice is both valid and impactful when used in authentic educational settings.

Looking ahead, the continued refinement of AQG pipelines, coupled with thoughtful integration of LLM-based personalized feedback and stronger faculty engagement strategies, points toward a future in which textbooks function as interactive, learning-by-doing environments that reliably maximize student learning potential.

References

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

[2] 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

[3] Koedinger, K. R., Kim, J., Jia, J., McLaughlin, E., & Bier, N. (2015, March). 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 (L@S'15),* pp. 111–120. https://doi.org/10.1145/2724660.2724681

[4] Koedinger, K. R., McLaughlin, E. A., Jia, J. Z., & Bier, N. L. (2016, April). Is the doer effect a causal relationship? How can we tell and why it's important. *Proceedings of the Sixth International Learning Analytics & Knowledge Conference (LAK '16),* pp. 388–397. http://dx.doi.org/10.1145/2883851.2883957

[5] Koedinger, K. R., Scheines, R., & Schaldenbrand, P. (2018). Is the doer effect robust across multiple data sets? In *Proceedings of the 11th International Conference on Educational Data Mining* (pp. 369–375).

[6] Van Campenhout, R., Johnson, B. G., & Olsen, J. A. (2021, July 18–22). The doer effect: Replicating findings that doing causes learning. In *The Thirteenth International Conference on Mobile, Hybrid, and On-line Learning* (*eLmL 2021)* (pp. 1–6). IARIA. https://www.thinkmind.org/index.php?view=article&articleid=elml_2021_1_10_58001

[7] Van Campenhout, R., Jerome, B., & Johnson, B. G. (2023). The doer effect at scale: Investigating correlation and causation across seven courses. *Proceedings of the 13th International Learning Analytics & Knowledge Conference (LAK '23),* pp. 357–365. https://doi.org/10.1145/3576050.3576103

[8] Van Campenhout, R., Dittel, J. S., Jerome, B., & Johnson, B. G. (2021). Transforming textbooks into learning by doing environments: An evaluation of textbook-based automatic question generation. In *Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education CEUR Workshop Proceedings*, (pp. 1–12). https://ceur-ws.org/Vol-2895/paper06.pdf

[9] 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 (L@S'22),* pp. 325–329. https://doi.org/10.1145/3491140.3528323

[10] Van Campenhout, R., & Hubertz, M. (2023). Context and considerations for investigating the impact of learning by doing on student equity. *Workshop on Equity, Diversity, and Inclusion in Educational Technology, Artificial Intelligence in Education (AIED)* (pp. 1–5). https://doi.org/10.5281/zenodo.8208452

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

[12] Van Campenhout, R., Johnson, B. G., Clark, M., Deininger, M., Harper, S., Odenweller, K., & Wilgenbusch, E. (2025). AI-generated questions in context: A contextualized investigation using platform data, student feedback, and faculty observations. *Journal of Communications Software and Systems, 21*(2), 178–188. https://doi.org/10.24138/jcomss-2024-0120

[13] 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

[14] 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

[15] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* pp. 404–411. https://aclanthology.org/W04-3252

[16] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *International Conference on Learning Representations (ICLR) 2013. Workshop proceedings.* https://doi.org/10.48550/arXiv.1301.3781

[17] Jerome, B., Van Campenhout, R., Dittel, J. S., Benton, R., & Johnson, B. G. (2023). Iterative improvement of automatically generated practice with the Content Improvement Service. In R. Sottilare & J. Schwarz (Eds.), *Adaptive Instructional Systems. HCII 2023* (Lecture Notes in

Computer Science, pp. 312–324). Springer. https://doi.org/10.1007/978-3-031-34735-1_22

[18] Dittel, J. S., Van Campenhout, R., & Johnson, B. G. (2025). Refining sentence selection for automatic cloze question generation with large language models. *The Twelfth ACM Conference on Learning @ Scale (L@S'25)*. https://doi.org/10.1145/3698205.3733926

[19] Van Campenhout, R., Dittel, J. S., & Johnson, B. G. (2026). Scaling effective characteristics of ITSs: A preliminary analysis of LLM-based personalized feedback. In S. Graf & A. Markos (Eds.), *Generative systems and intelligent tutoring systems. ITS 2025. Lecture Notes in Computer Science* (Vol. 15723). Springer. https://doi.org/10.1007/978-3-031-98281-1_13

[20] 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

[21] Kulik, J. A., & Fletcher, J. D. (2016).Effectiveness of intelligent tutoring systems:A meta-analytic review. *Review of Educational Research, 86*(1), 42-78.

[22] 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). Longman.

[23] Goldstein, P. J., & Katz, R. N. (2005). Academic analytics: The uses of management information and technology in higher education. Educause. https://library.educause.edu/-/media/files/library/2005/12/ers0508w-pdf.pdf

[24] 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

[25] 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

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

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

[28] 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, pp. 286–296). Springer, https://doi.org/10.1007/978-3-031-22131-6_22

[29] Van Campenhout, R., Clark, M., Johnson, B. G., Deininger, M., Harper, S., Odenweller, K., & Wilgenbusch, E. (2024). Automatically generated practice in the classroom: Exploring performance and impact across courses. In *Proceedings of the 32nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2024)* (pp. 1–6). https://doi.org/10.23919/SoftCOM62040.2024.10721828

[30] Van Campenhout, R., Autry, K., Clark, M. W., & Johnson, B. G. (2025). Scaling the doer effect: A replication analysis using AI-generated questions. *Proceedings of the Twelfth ACM Conference on Learning@Scale (L@S'25)*. https://doi.org/10.1145/3698205.3729545