Balancing Efficacy and Engagement in Interactive Texts

Jonah Bard[1]

[1]Dartmouth College, Hanover, NH 03755, USA

Abstract

We introduce Phosphor, a digital learning platform that integrates LLM-graded formative assessment directly into instructional content, and report results from a deployment with 151 students across three sections of Introductory Statistics at Dartmouth College. Full dosage of the Phosphor material is associated with an increase in final exam performance of between 0.71 SD (adjusting for prior exam scores) and 1.30 SD (unadjusted). Presented as an entirely optional, ungraded alternative to traditional readings, the platform was adopted by 90.2% of enrolled students, far exceeding typical reading-compliance rates. Additionally, score results across a natural variation in quiz formats suggest that embedded constructed-response questions are a valuable ingredient in driving outcomes. Together these results indicate that high engagement and measurable efficacy are simultaneously achievable; we discuss implications for the design of AI-augmented instructional tools.

Keywords

intelligent textbooks, large language models, formative assessment, intelligent tutoring, retrieval-augmented generation, reading engagement, constructed-response assessment

1. Introduction

Now is the best time in history to be working on personalized instructional software. With the advent of LLMs, every student has a 24/7 knowledgeable personal tutor in their pocket. But if a basic LLM is all students need, why isn't academic achievement suddenly skyrocketing?

It appears that unrestricted, external AI use is largely a hindrance to students despite its convenience. Bastani et al. [1] demonstrated in a randomized controlled trial with nearly 1,000 students that unfettered access to GPT-4 actually harmed subsequent performance by 17% when the tool was removed — students used it as a crutch rather than a learning aid. Only a version with carefully designed pedagogical guardrails mitigated these negative effects. Meanwhile, student use of generative AI for academic work has surged: a 2026 survey by the Higher Education Policy Institute found that 94% of university students reported using generative AI on assessed work, up from 53% just two years earlier [2].

A recent meta-analysis suggests an overall positive effect of LLMs for targeted use cases [3], though broadly effective interventions proven in rigorous experimental designs remain scarce.

Against this backdrop, it is well documented that university students almost never read the textbook. Reading compliance among undergraduates has declined dramatically since the 1980s [4], with students actively resisting reading assignments, and objective measures consistently showing substantially lower compliance than students self-report [5]. In our deployment, student-reported reading completion baselines for MATH 010 were approximately 15%, with instructors estimating 10%. Individual student reports of reading compliance ranged from "literally no one does that" to "is this being recorded?"

These observations motivated the design of Phosphor (f.k.a. Spongium), a digital learning platform integrating LLM-powered formative assessment into instructional content. Phosphor embeds AI-graded quizzes into the reading workflow, making active recall a structural feature of the learning experience. The core premise is that AI is most effective when integrated directly into the content delivery system — a design philosophy aligned with the emerging vision of intelligent textbooks [6] and supported by the well-documented "doer effect," in which completing practice questions integrated into readings yields several times the learning impact of reading alone [7, 8].

iTextbooks'26: Seventh Workshop on Intelligent Textbooks, June 28, 2026, Seoul, Republic of Korea
jonah.z.bard.27@dartmouth.edu (J. Bard)
https://jonahbard.com (J. Bard)

2. Platform Design and Deployment

2.1. Platform Overview

Phosphor is deployed as a web application. Instructional content is organized into lessons, each rendered as navigable web pages with a sidebar showing the full curriculum and per-lesson completion indicators. The statistics curriculum was custom-authored, grounded in open educational resources. The platform includes the following:

Lesson Quizzes. Each lesson is associated with a bank of 15–20 exercises. Students take Lesson Quizzes consisting of four randomly selected questions from the lesson's bank. Multiple-choice questions (MCQ) are auto-graded; constructed-response questions (CRQ) are graded by Claude Sonnet 4.6 against instructor-defined, question-specific rubric criteria. The grading prompt receives the student's response alongside the question text, a model answer, and explicit grading criteria, and returns a correctness judgment with an explanation. Students who achieve 75% or higher are marked as having "passed" and have "completed" the quiz. The test bank consists of 40% CRQ and 60% MCQ. Content is not gated: students may freely read and take quizzes regardless of past results. Quizzes permit unlimited retries.

The use of LLMs for constructed-response grading has shown promising alignment with human raters [9], though reliability varies with question complexity. We did not conduct a formal inter-rater reliability study; however, the grading prompt followed best practices in LLM-based assessment [10].

Module Reviews. Phosphor offers cumulative Module Review quizzes covering content across all lessons in a module—each containing 10 questions, with a 90% pass threshold. The Module Review is MCQ-only by default but features an "all question types" mode in which students can opt to include a combination of CRQ and MCQ in the quiz. Students have unlimited retries.

RAG-based chat assistant. A retrieval-augmented generation (RAG) chat sidebar allows students to ask questions while reading. The student's query is embedded and matched against a vector index of curriculum content via cosine similarity, with top-matching chunks placed into the LLM's context alongside guardrails restricting responses to the boundaries of the course curriculum.

2.2. Deployment Context

Phosphor was deployed in an early pilot across three sections of MATH 010 (Introductory Statistics) at Dartmouth College in Spring 2026. Enrollment followed typical withdrawal rates, starting at 151 students and finishing at 143. Designed for non-math majors, MATH 010 is typically taken by underclassmen. The platform was presented as an entirely optional, ungraded alternative to traditional course readings.

The course curriculum was organized into three modules, aligned to two Midterm exams and one cumulative Final Exam. We evaluate results from each exam, which were administered on-paper and in-person with heavy proctoring. As Phosphor was deployed within regular course instruction, the results reported here draw on a secondary analysis of de-identified, aggregate student records.

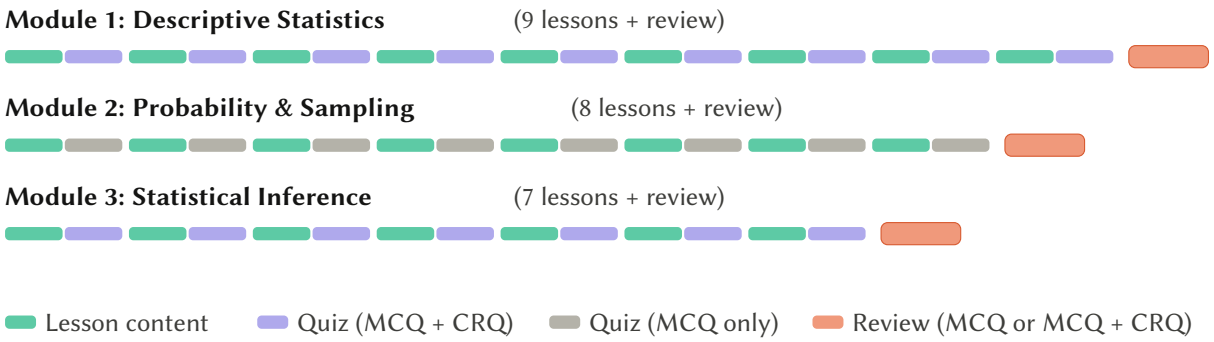
Critically, the quiz format differed between modules due to iterative design changes made in response to student feedback:

Module 1 (Descriptive Statistics, 9 lessons). Each lesson included a mixed-format quiz containing both MCQ and CRQ, all graded by the LLM against rubric criteria.

Module 2 (Probability & Sampling, 8 lessons). In response to widespread student feedback that the CRQ auto-grader was rigid and discouraging, Lesson Quizzes were reduced to MCQ only. The Module Review was also introduced in response to strong student demand.

Module 3 (Statistical Inference, 7 lessons). After analyzing exam results, which suggested that MCQ-only Lesson Quizzes provide negligible learning benefits, CRQ were re-introduced to them. The Module Review was available in Module 3 as well.

Note: the Module 1 Review was introduced after Midterm 2, upon request from students who desired more formative assessment resources when studying for the cumulative Final Exam.



3. Results

3.1. Engagement

Table 1 reports cumulative student engagement at various thresholds. 90.2% of enrolled students engaged with the platform, via Module Reviews or Lesson Quizzes, at least once. The engagement threshold table below uses the final denominator of 143 enrolled students, among which the median lesson reached was 22 of 24 (91.7%). Specifically among the 137 students who created a Phosphor account, the median lessons reached was 23 of 24 (96%).

We assemble multiple measures of total reading compliance. An upper bound is given by exposure, with 75.6% of student–lesson pairs reached (taking the later of each student's last-viewed and furthest completed lesson). The aggregate quiz completion rate was 48.1%. Because completing a lesson's quiz requires having read the lesson, this is interpreted as a lower bound. Total reading compliance thus falls within $[48\%, 76\%]$, against student- and instructor-reported baselines of 10–15% for this course.

Table 1 also reports survey results administered live, in-class to each of two course sections via Mentimeter, once immediately after Midterm 1 (S1, $n = 33$) and once immediately after Midterm 2 (S2, $n = 31$); we report these as descriptive indicators given potential social desirability bias.

Table 1
Platform engagement and student reception. *Compl.* = lesson quiz completed (a lower bound on reading, since passing assumes having read the lesson); *Reached* = furthest lesson reached, via last-viewed page or furthest completion (an upper bound, assuming sequential progression). The threshold panel is over the final roster ($N = 143$); the per-module panel uses per-module exam-taker denominators. "Passing" a Module Review consists of achieving a minimum 90% score.

Overall Completion Metrics

Threshold	Compl.	Reached
≥5 lessons	69.9%	87.4%
≥10 lessons	57.3%	81.8%
≥15 lessons	39.9%	76.2%
≥20 lessons	25.2%	59.4%
All 24 lessons	11.2%	44.8%
Median	45.8%	91.7%
Mean	48.0%	75.6%
≥1 Review	76.9%	—
≥1 LQ or MR	90.2%	—

Lesson Quiz Engagement

Module	Compl.	Rate
M1	670/1350	49.6%
M2	691/1176	58.8%
M3	334/1001	33.4%
All	1695/3527	48.1%

Module Review Engagement

Mod.	n	Took	Pass
M1	150	48.0%	34.7%
M2	147	70.1%	61.2%
M3	143	50.3%	36.4%
All	440	56.1%	44.1%

Survey Results (% agree)

Metric	S1	S2
More engaging	94%	94%
Better retention	94%	97%
Prepared for class	76%	87%
Assess accurately	79%	–
Adequate for exam	79%	61%
Other courses would benefit	85%	90%
n	33	31

3.2. Learning Outcomes

For each module we examine the relationship between platform dosage and exam performance. For the two midterms we use the corresponding module's lessons; for the cumulative Final Exam we use completions across all 24 lessons. The Final contained 24 graded questions, 8 per module, plus one universal free point; we analyze the 24-point graded score. Group comparisons use Welch's unequal-variance t-test (one-sided); effect sizes are Cohen's d with pooled SD. We treat the five pre-specified binary contrasts in Table 3 as a single family and report Holm-adjusted significance (family-wise $\alpha = 0.05$); the correction is scoped to these contrasts, with dosage regressions (Table 2) and Tobit models (Table 4) treated as estimation. Of 151 students enrolled, 150 took Midterm 1, 147 took Midterm 2, and 143 took the Final; dosage and Tobit analyses use the 138 students with records for each exam.

Table 2
Linear regressions of exam grade (0–100) on lesson completions among platform users. Module 1 quizzes included CRQ; Module 2 quizzes were MCQ-only; Module 3 restored CRQ. The Final Exam's regression uses completions across all 24 lessons.

	Module 1 → Midterm 1	Module 2 → Midterm 2	All lessons → Final
All platform users	$1.64x + 77.9$, $R^2 = 0.123$	$0.90x + 76.0$, $R^2 = 0.027$	$0.41x + 84.7$, $R^2 = 0.091$
$\geq$1 lesson completion	$1.37x + 80.0$, $R^2 = 0.111$	$-0.29x + 84.7$, $R^2 = 0.001$	$0.45x + 84.0$, $R^2 = 0.096$

Dosage–Performance Regressions. Table 2 reports these regressions. For Module 1, each additional lesson completion is associated with roughly 1.6 additional percentage points on the first midterm ($p < 0.001$, $R^2 = 0.123$), and the association holds whether or not zero-completion users are included. For Module 2, the apparent positive slope among all users ($R^2 = 0.027$) reflects only the zero-versus-nonzero distinction; among students with at least one completion the slope is slightly negative ($R^2 = 0.001$), indicating no dosage relationship for MCQ-only quizzes. On the cumulative Final, each completion is associated with about 0.4 additional points ($R^2 = 0.091$), and—unlike Module 2—this is essentially unchanged when zero-completion students are excluded ($R^2 = 0.096$).

Table 3
Welch's t-tests comparing exam performance by platform engagement. One-sided p-values test the directional hypothesis that engaged students outperform non-engaged. Welch–Satterthwaite df in the df column. d is Cohen's d (pooled). Final-Exam contrasts use the 24-point graded score.

Comparison ($n_{\text{eng}}/n_{\text{non}}$)	t	df	Δ	p	d
Midterm outcomes					
M2: Completed review → MT2 (80/67)	2.17	132	+5.7	0.016	0.36
M2: Passed review → MT2 (72/75)	2.38	144	+6.1	0.009	0.39
Final-Exam outcomes (cumulative)					
$\geq$1 lesson completion → Final (118/20)	1.31	24	+3.9	0.101	0.36
$\geq$1 review pass → Final (96/42)	1.75	60	+4.1	0.042	0.37
All 3 reviews passed → Final (31/107)*	4.27	85	+7.1	<0.0001	0.66

*Survives Holm correction across the five pre-specified binary contrasts (family-wise $\alpha = 0.05$).

Binary Comparisons. On the Final Exam, the per-lesson contrast between Phosphor users and non-users is positive (+3.9 points, $d = 0.36$, $p = 0.101$), while Module Reviews give the strongest signal: students who passed all three Module Reviews scored 7.1 points higher ($d = 0.66$, $p < 0.0001$). Passing at least one review is also positively associated with the Final (+4.1 points, $d = 0.37$, $p = 0.042$). The Module 2 review corroborates this internally: passing was associated with a 6.1-point Midterm 2 advantage (nominal $p = 0.009$, $d = 0.39$), holding content, timing, and cohort fixed—the same direction and comparable magnitude.

Cumulative Engagement and Selection. To summarize the two engagement channels jointly, we fit a Tobit model of the Final Exam (Table 4), right-censored at the 24-point ceiling reached by 27% of students. Under the engagement-only specification, the gap between full engagement (24 lessons, 3 reviews) and zero engagement is 14.7 points on a 0–100 scale (1.30 SD), with both coefficients positive though individually modest ($z \approx 1.8$). Controlling for student performance in either midterm roughly halves the gap to 8.0 points (0.71 SD), absorbing much of the review coefficient ($z \approx 0.6$) while leaving the per-lesson coefficient intact ($z \approx 1.8$); the review benefit appears already banked by midterm time, whereas dosage carries independent end-of-term signal.

Table 4
Joint Tobit models of Final Exam score, right-censored at 24-point ceiling accommodating the 27% of students at maximum. Coefficients on latent scale; z in parentheses. Controls enter as corresponding midterm score.

	(1) Engagement only	(2) + Midterm 1	(3) + Midterm 2
Lesson completions (0–24)	0.074 (1.8)	0.060 (1.8)	0.058 (1.6)
Reviews passed (0–3)	0.584 (1.8)	0.162 (0.6)	0.180 (0.7)
Midterm control	—	15.55 (8.0)	11.63 (6.9)
Intercept (/24)	20.49	7.38	11.54
σ (/24)	3.17	2.50	2.65
N (censored at ceiling)	138 (37, 27%)	138 (37, 27%)	136 (37, 27%)
Full-vs-zero gap (/100)	**14.7**	**8.0**	**8.0**
Full-vs-zero gap (SD)	**1.30**	**0.71**	**0.71**

Gap = predicted latent Final for full engagement (24 lessons, 3 reviews) minus zero engagement; expressed as points on a 0–100 scale (raw/24×100) and in units of the observed Final SD (2.72 points on the 24-point scale). Midterm coefficients are on each midterm's raw scale and are not comparable across columns 2–3. Column 3 uses $N = 136$ (two students lack a usable Midterm 2); the engagement-only gap on that subsample is 3.52.

Module Review Retry Behavior. Across all three Module Reviews, two-thirds of student attempts (157 of 237) involved returning for at least one retry, and these retries were overwhelmingly spaced rather than immediate—only 29% came back within an hour while 55% returned a day or more later (median gap ∼1.5 days). Retry rates were consistent across modules (M1 61%, M2 77%, M3 56%).

3.3. RAG Chat Underutilization

The RAG chat assistant saw relatively minimal usage, with 72 total queries, and only 14 students submitting more than one. Students reported two reasons: that general-purpose LLMs were faster and more capable for their questions, and that the reference content was "sufficient" such that they did not generate enough questions during reading to justify a separate chat interface.

4. Discussion

Taken together, the results point to a substantial cumulative benefit from full engagement with Phosphor. Jointly modeling the two engagement channels (Lesson Quiz completions and Module Reviews), the Tobit estimates place the gap between full and zero engagement at 1.30 SD on the Final Exam before adjustment. Because the most engaged students are also the most able and motivated, we control directly for prior achievement by conditioning on midterm performance, which attenuates the gap to 0.71 SD. We read 0.71 SD as a conservative lower bound rather than a point estimate: the cumulative Final re-tests content already assessed at the midterms, so the midterm control is largely a parallel measure of the outcome, absorbing learning that Phosphor itself may have produced earlier in the term. The defensible cumulative effect therefore lies between roughly 0.71 SD (over-adjusted) and 1.30 SD (selection-inflated)—large by observational standards for educational interventions.

A strong point of evidence that format, rather than motivation alone, is a primary driver of the effect is the result of a natural variation in quiz design between modules. Lesson-level dosage tracked exam

performance under constructed-response quizzes but not under multiple-choice quizzes (Table 2), despite comparable-or-higher engagement. When varying the assessment format, engagement translated into measurable learning only where the format demanded active generation. This is consistent with the testing-effect literature: Kang et al. [11] found that short-answer quizzes with feedback produced stronger retention than multiple-choice quizzes ($d = 0.41$). The cumulative Final corroborates the pattern at the whole-course level—each lesson completion is associated with roughly 0.4 additional points, essentially unchanged when zero-completion students are excluded ($R^2 = 0.091 \rightarrow 0.096$)—indicating that the cumulative signal is carried by the CRQ lessons of Modules 1 and 3 rather than by the engaged/non-engaged boundary alone.

Within this picture, cumulative Module Reviews emerge as the strongest single lever. Students who passed all three Reviews scored 7.1 points higher on the Final ($d = 0.66$)—the largest effect in the study. Because that group is also the most self-selected, we anchor the claim on the cleaner within-module comparison: in Module 2, passing the review predicted a 6.1-point Midterm 2 advantage ($d = 0.39$). The Module Reviews differed from Lesson Quizzes along several dimensions (cumulative scope, interleaved topics, higher threshold, optional CRQ), so we cannot isolate the operative factor; the most likely contributor is interleaved retrieval practice across topics [12], compounded by the spaced-review behavior suggested by the retry data.

The minimal usage of the RAG chat assistant is consistent with emerging patterns—Khan Academy recently reported that only 15% of users regularly engage with their supplementary chatbot [13]—and suggests that integrating targeted AI features into content delivery through assessment, feedback, and progress tracking may be a more immediately productive design.

These findings suggest several design principles for AI-augmented textbook systems. Crucially, LLMs make it feasible to grade formative CRQ against rubric criteria at scale, a capability that appears pedagogically significant rather than merely convenient. Notably, strong student satisfaction and voluntary engagement mostly coexisted with CRQ: our first survey was administered before the transition to MCQ-only Lesson Quizzes, so the format that best predicted learning was not rejected wholesale. We do note, however, that disabling CRQ for Module 2 was followed by a modest jump in lesson completions, while re-enabling them was followed by a decline (Table 1). Thus, there may still emerge an engagement tradeoff for the highest-efficacy interventions.

We emphasize that as base LLMs become more powerful, their capacity for distraction to students will only rise. This study indicates that it is indeed possible to create highly effective learning experiences that are engaging enough such that students will opt to use them rather than rely on external AI tools. Further, we highlight that the strong benefits of AI-powered personalized instruction can be successfully deployed in an institutional setting.

Limitations. This is an observational study of a pilot deployment, at a single selective institution, and lacks randomized controls. Self-selection is the central threat: students who complete more quizzes may be more motivated or higher-performing generally. Our joint Tobit models control for prior achievement, generating a lower- and upper- bound effect size estimate. However, the cross-module lesson contrast remains confounded by content domain, timing, and the simultaneous introduction of the Module Review, and the all-reviews-passed group is the most self-selected in the study. Survey data were collected from small samples with potential social desirability bias.

Future Work. A controlled study with random assignment to quiz format conditions would strengthen the causal claim about CRQ versus MCQ. We plan to deploy Phosphor with a completion requirement attached to course grade, which we anticipate will increase engagement and provide a cleaner dosage–performance measure. Replication in additional university gateway courses and across institutions is a strong priority.

Acknowledgments

We thank Thomas Zdyrski, Keenan Eikenberry, Erik van Erp, Christophe Hauser, Katherine Salesin, and J. Peter Brady for their generous support in the development and rollout of Phosphor.

Statement on Generative Artificial Intelligence

In compliance with CEUR guidelines, we describe the ways in which generative AI was used in this work: Claude was utilized for background research, copyediting, and proofreading. All outputs were reviewed and edited meticulously by the author, who takes full responsibility for the publication's content.

References

[1] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, R. Mariman, Generative AI without guardrails can harm learning: Evidence from high school mathematics, Proceedings of the National Academy of Sciences 122 (2025) e2422633122.

[2] R. Stephenson, C. Armstrong, Student generative AI survey 2026, HEPI Policy Note, 2026. URL: https://www.hepi.ac.uk/reports/student-generative-ai-survey-2026/.

[3] Y. Ma, C. Zhong, A meta-analysis of the impact of generative artificial intelligence on learning outcomes, Journal of Computer Assisted Learning 41 (2025) e70117. doi:10.1111/jcal.70117.

[4] C. M. Burchfield, J. Sappington, Compliance with required reading assignments, Teaching of Psychology 27 (2000) 58–60.

[5] J. Sappington, K. Kinsey, K. Munsayac, Two studies of reading compliance among college students, Teaching of Psychology 29 (2002) 272–274.

[6] P. Brusilovsky, S. Sosnovsky, K. Thaker, The return of intelligent textbooks, AI Magazine 43 (2022) 337–340. doi:10.1002/aaai.12061.

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

[8] R. Van Campenhout, B. G. Johnson, J. A. Olsen, The doer effect at scale: Investigating correlation and causation across seven courses, in: Proceedings of the 13th International Learning Analytics and Knowledge Conference (LAK '23), ACM, 2023, pp. 357–365.

[9] O. Henkel, L. Hills, B. Roberts, J. McGrane, Can large language models make the grade? An empirical study evaluating LLMs' ability to mark short answer questions in K-12 education, in: Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24), 2024. doi:10.1145/3657604.3664643.

[10] E. Latif, X. Zhai, Fine-tuning ChatGPT for automatic scoring, Computers and Education: Artificial Intelligence 6 (2024) 100210. doi:10.1016/j.caeai.2024.100210.

[11] S. H. K. Kang, K. B. McDermott, H. L. Roediger, III, Test format and corrective feedback modify the effect of testing on long-term retention, European Journal of Cognitive Psychology 19 (2007) 528–558. doi:10.1080/09541440601056620.

[12] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, D. T. Willingham, Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology, Psychological Science in the Public Interest 14 (2013) 4–58.

[13] K. E. DiCerbo, Kristen's corner winter 2026, Khan Academy Blog, 2026. URL: https://blog.khanacademy.org/kristens-corner-winter-2026/.