A/B Testing on Open Textbooks

A Feasibility Study for Continuously Improving Open Educational Resources

Royce Kimmons

This study examined the feasibility of employing A/B tests for continuous improvement by focusing on user perceptions of quality of six chapters of a popular open textbook over the course of a year. Results indicated non-significant differences in all cases but also suggest that future work in this area should (a) employ A/B testing at a broader, less-granular (e.g., platform-level) scale to increase sample sizes, (b) explore autonomous approaches to experimentation and improvement, such as bandit algorithms, and (c) rely upon more universally collected dependent variables to reduce sample size limitations emerging from self-reports.

Open educational resources provide great promise to instructional designers as low-cost, high-impact educational materials that can be used, shared, remixed, and adapted with ease. Especially when viewed through the lens of the "5Rs" of openness (Wiley, n.d.)—Retain, Revise, Remix, Reuse, Redistribute—or the lens of "expansive openness" (Kimmons, 2016), such resources give instructional designers the ability to create and share learning materials at a massive scale, to adapt existing resources for better meeting the needs of target learners, and to remix resources from

various authors into multi-faceted and rich learning experiences.

Because of the ubiquity of textbooks in higher education, the open textbook as a medium promises to be a valuable means for providing learning opportunities to many students while also driving down costs. Students at four-year universities in the U.S. currently spend an average of $1,240 on textbooks per year (College Board, 2019), and textbook cost hikes have far outpaced inflation, consumer costs, and recreational book costs, making higher education opportunities more cost-prohibitive and requiring students to skip meals, enroll in fewer courses, and work longer hours (Whitford, 2018). While open textbooks provide an opportunity for universities to drive down student costs and to improve learning experiences, open textbooks are not widely used (Seaman & Seaman, 2018). This is presumably due to perceptions of time limitations emerging from tenure and promotion practices and perceptions that open textbooks are of relatively poor quality when compared to their copyright-restricted alternatives (Kimmons, 2015; Martin & Kimmons, 2020).

Though systemic challenges to open textbook adoption may be outside the realm of instructional designers to address, one clear way that we can make a difference is to help improve the quality of these resources. Some initial work has sought to establish quality metrics for open textbooks and other open resources (Bodily et al., 2017; Woodward et al., 2017), and Dinevski (2008) proposes that the quality control of these resources is relatively unique by placing accountability in the hands of learners, teachers, and local designers to address localized or demographic-specific needs, rather than upon market-driven publisher considerations. Furthermore, though traditionally published textbook editions are viewed as static entities that are either high- or low-quality, because of their live and open nature, open textbooks can also undergo continuous improvement efforts that iteratively improve their quality over time, correcting mistakes, refining formatting, and providing supplements as needed to improve learning (Wiley et al., 2021).

For these reasons, applying continuous improvement cycles to open educational resources is of increasing interest to designers, but we are only just beginning to figure out how to do this well, especially when large-scale data are involved and resources are being used by a wide array of learners. Borrowing from the software development field (the same field where the notion of openness came from, to begin with; Kimmons, 2016; Open Source Initiative, n.d.; Stallman, 2013), it seems reasonable to consider how modern approaches to software improvement might apply to educational resources as well. As a promising example, A/B or split testing is an approach to software development that places at least two different versions of a product in front of random sets of actual users and analyzes their behaviors over time to determine which is superior (Kohavi & Longbotham, 2017).

When it comes to education, A/B testing has been proposed not only as a process for improving design but also as a process for choosing between competing pedagogical methods or other decisions of educational importance (UpGrade, n.d.). In the case of open textbooks, A/B testing would require having at least two versions of content that users interact with. The "A" version (otherwise called the original version or control) represents the default version of the resource as originally created by the author, while the "B" version (otherwise called the experimental flight or fork) represents a variation of the resource that the researcher hypothesizes might yield differing behaviors or results. To make comparisons, audience size for each version may not need to be equal, and relative sampling for different versions may involve an assessment of the urgency and relative importance of experimental variations. As readers are assigned to the competing versions of the textbook, a variety of analytics could be collected to test which version is superior, and successive tests could theoretically be employed on the same resource to gradually improve it in many different ways.

Bringing these ideas together, this study explores the feasibility of using A/B testing to inform continuous improvement and increase the

perceived quality of open textbooks. Relying upon data collection and analysis mechanisms of a popular open textbook for undergraduate and teacher education, the guiding research question of this study was “How feasible is it to conduct A/B testing on highly-used open textbook chapters for the purpose of improving perceptions of quality?”

Methods

To conduct this study, experimental flights were created within the EdTech Books system by copying six chapters as new flights (or “B” versions), adjusting their contents, and setting each chapter’s “Flight Mode” to “Automatic.” The automatic mode meant that whenever any reader navigated to the chapter, they were randomly assigned to either view the original or the experimental flight. This assignment was done without the reader’s awareness and ensured true randomization. Flight assignment was enabled for a period of 12 months (February 2020 to February 2021), and results were then analyzed to compare reader behaviors and perceptions for the time period. As a methodological note, though this timeframe coincided with the COVID-19 pandemic in many countries and resulting shifts to online and remote learning might have influenced overall usage of open resources, such a shift would not be expected to influence the types of user behaviors measured here between groups. For instance, though more people might have started reading the textbooks because of the pandemic, we would not expect this to influence the relationship between text size within the textbooks and reading behaviors. For this reason, we did not conclude that the targeted timeframe for the study should be considered as an additional variable or meaningful frame of analysis.

Context

EdTech Books is a free online publishing platform for open textbooks.

Built with PHP, MySQL, and Javascript, the platform operates on four guiding values of freedom, accessibility, usability, and quality, providing authors with tools to easily create, remix, and share textbooks (Kimmons, n.d.). Currently, the platform provides content to roughly 50,000 unique readers per month, representing students, teachers, and the general public. Content is provided in simple HTML via web pages and also as PDFs for download, representing millions of page views over the course of its two-year lifespan.

Central to the mission and design of EdTech Books is the goal of supporting continuous improvement and improved perceptions of open textbook quality. Toward this end, the system provides A/B testing features, quality assurance mechanisms, advanced analytics, and various other tools to support ongoing analysis, adjustment, and improvement of materials. However, since the notion of continuous improvement is not commonly connected to the development of published materials, like textbooks, it is unclear how to do this well and how to develop systems that both empower and encourage authors to engage in this process.

For this study, I analyzed results from six experiments conducted within EdTech Books upon separate chapters of a popular open textbook: *The K-12 Educational Technology Handbook* by Ottenbreit-Leftwich and Kimmons (2020). This textbook has been accessed over 120,000 times in its short lifespan and is widely used for teacher education courses and professional development efforts and is also commonly accessed from search engine results on topics related to technology's role in education.

Participants

As readers accessed the textbook on the platform for the first time, they were notified that the system collects anonymous analytics related to their behaviors, and they were given the option to opt-out of being tracked in this way. For this study, I focused on opted-in reader

data associated with this single textbook.

As with other textbooks in the platform, readers of the textbook accessed chapters in many ways but generally fell into two categories: (a) formal learners who accessed chapters from links or LMS embeds associated with official university courses and (b) non-formal or informal learners who accessed chapters from organic search engine results (e.g., those searching Google for “tech integration”). Backlink analysis of the textbook revealed that it was heavily used by students at a number of universities, including Brigham Young University, Marist University, Oklahoma State University, State University of New York, Montana State University, Purdue University, and others. The breakdown of formal vs. non/informal learners, however, varied from chapter to chapter with some chapters like “Technology Integration” experiencing a relatively even split between the two and others exhibiting high skew in one direction or the other. Even within these categories, we would expect to find great variation in reader goals, purposes, and activities, as higher education institutions use these resources for diverse courses. For the purpose of this study, reader type was not considered in data analysis, and the flight assignment procedure did not take reader category into consideration for random assignment, meaning that the demographics of both the original and experimental versions of each chapter would be expected to exhibit similar distributions of reader types to the overall chapter. This was an intentional design decision but assumes that optimal design decisions for improving perceived quality would not vary by reader category.

Dependent Variable

Because perceptions of poor quality are a major barrier to open textbook adoption and diffusion (Kimmons, 2016; Martin & Kimmons, 2020) and the improvement of perceived quality is a major goal stated on the platform, we constructed experiments with the goal of improving reader perceptions of quality, as measured by a simple

survey. This single-question survey was provided as an unobtrusive "End-of-Chapter Survey" at the bottom of each chapter that asked the following: "Overall Quality: How would you rate the overall quality of this chapter?" Possible responses were coded to an ordinal scale as follows: (1) "Very Low Quality," (2) "Low Quality," (3) "Moderate Quality," (4) "High Quality," and (5) "Very High Quality." The form was then automatically submitted as readers navigated away from the chapter or closed their browser tab, resulting in an average quality rating of 4.1/5.0 for the targeted textbook chapters (n = 963 ratings, SD = .67). Results also exhibited a strongly negative skew, with only 4 ratings (0.4%) falling below "Moderate Quality" (see Figure 1). These ratings represented results from 810 different users with the average user leaving 1.19 ratings across chapters in the book (SD = .75, Max = 10).

Figure 1

Distribution of Textbook Ratings

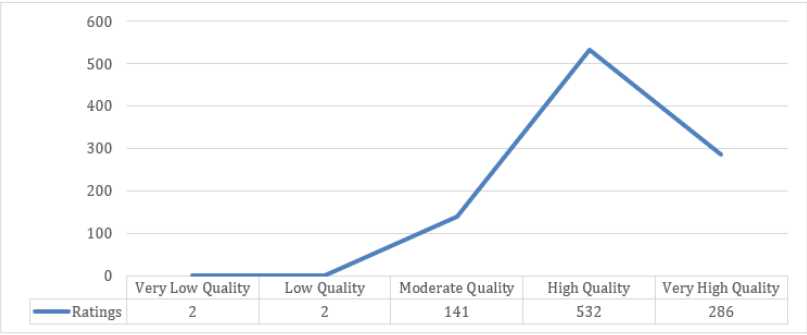


Chart showing the distribution of textbook ratings

The unobtrusive and optional nature of this survey helped to avoid Hawthorne effects in results and provided similar benefits to those found in the analysis of public internet data sources (Kimmons &

Veletsianos, 2018), even though some interpretive power was lost with limited contextual information about readers. This approach also provided minimal risk, effort, and discomfort to users and prevented analyses from being classified as human subjects research according to NIH definitions, because the process (a) did not collect information about individuals and (b) did not include identifiable data, such as demographics, names, user type information (e.g., student vs. faculty), or IP addresses. This means that the sample size for each experiment was limited to those who anonymously answered the quality assurance measure at the end of the chapter, which accounted for around 1% of readers for each chapter.

Though such a low response rate would be troubling in some research settings, the fact that readers were randomly assigned to the two groups helps to alleviate concerns of self-selection bias, and low rates of response will always be a necessity when using unobtrusive measures of relatively free-roaming user activities like these. This point is of special importance when studying open resources, because most of the traffic (or user behavior) associated with these resources constitutes lurking (Bozkurt et al., 2020) or those who may briefly open the chapter without any intent to actually read it. To illustrate, Google Analytics reported that the bounce rate for the book in this time period (or the number of users who navigated away after viewing only one page) was 71.85% with the average user session lasting less than 3 minutes. This is why, for instance, MOOCs have such notoriously low completion rates (Gütl et al., 2014; Rivard, 2013) and why when studying open environments and resources it makes sense to limit analyses to users whose behaviors suggest an intent to participate in the behaviors we are measuring (e.g., Veletsianos et al., 2021). Judging by user scrolling behaviors, time on page, textual length, and chapter text complexity for the target textbook, it is estimated that only about 22.7% of page views actually constituted a “read” of the contents, and among those who read the contents, there was no incentive or prodding to complete the end-of-chapter survey. Yet, such data should nonetheless be valuable for understanding user

perceptions of resources in the same way that user ratings are valuable on sites like Amazon or Yelp to determine the quality of products or services, even if the relative representation of ratings is very small in comparison to the total number of customers on those sites.

Embedded automatically by the platform at the end of every chapter, quality assurance surveys provided results to authors in an "Analytics" dashboard at the flight, chapter, and book levels (see Figures 2 and 3). In the "Analytics" dashboard at the flight level, an additional table was also provided to authors that provides statistical comparisons between the original and the experimental flight (see Figure 4). These tables allowed authors to compare reader behaviors between the original and the experimental flight on the "Overall Quality" measure as well as embedded learning checks and surveys in the chapter. In the provided example, for instance, each row (except for the final "Overall Quality" row) represents a different learning check within the chapter, and the table reveals to the author whether the experimental flight influenced performance on the learning measure. Because these learning measures are chapter-dependent, they cannot be compared between chapters and will not be included in this study. However, common learning measures could be compared in future studies as readers are more likely to complete these than quality assurance surveys, thereby providing more robust sample sizes at a faster rate.

Figure 2

Screenshot of the Analytics Overview for a Chapter on EdTech Books

Overall Rating	4.1/5.0 ★★★★☆
Total Ratings	253
Page Views	19.0K
Tracked Views	20.7K
PDF Downloads	443
Cost Savings	$1.2K
Reading Ease	Very Difficult (28.1)
Grade Level	12+
Word Count	5,111
Reading Time	27 minutes
Predicted Reads	5.9K
Reading Likelihood	28%
Last Updated	2020-06-28 17:28:21

Chart showing the analytics categories to evaluate a chapter on EdTech Books

Figure 3

Screenshot of a Chapter Quality Display for a Chapter

Selection	Votes
Very Low Quality	0
Low Quality	1
Moderate Quality	47
High Quality	133
Very High Quality	72

Screenshot Showing the Chapter Quality Ratings

Figure 4

Screenshot of a Flight Comparison Table

	Original			no stock photos						
	Mean	n	SD	Mean	n	SD	Change	Welch t-Test	p-value	Cohen's d
teacher-values	1.37	336	0.82	1.46	270	0.85	0.09	1.32	NS	0.13
networked-thinking	0.72	410	0.69	0.78	308	0.55	0.06	1.35	NS	0.16
stimulus	0.81	389	0.58	0.78	299	0.65	-0.03	-0.69	NS	0.09
inner-workings	0.74	386	0.71	0.77	298	0.7	0.03	0.49	NS	0.05
prior-experiences	0.75	379	0.6	0.81	298	0.54	0.05	1.22	NS	0.17
tech-admin-values	1.6	331	0.68	1.64	261	0.65	0.05	0.89	NS	0.11
principal-values	0.96	332	0.84	0.97	259	0.83	0.01	0.12	NS	0.01
pck	0.69	327	0.67	0.63	256	0.81	-0.06	-0.99	NS	0.11
pic	0.81	331	0.65	0.85	255	0.63	0.04	0.76	NS	0.1
rat	0.94	333	0.36	0.91	256	0.43	-0.02	-0.68	NS	0.14
usefulness	3.93	336	0.96	3.89	266	0.93	-0.04	-0.49	NS	0.04
Overall Quality	4.09	256	0.7	4.19	195	0.63	0.11	1.66	NS	0.23

Chart showing a flight comparison table

Independent Variables

To improve perceived quality of the targeted chapters, format- and content-based experiments were created for six different chapters in the textbook, with each experimental flight representing a different variable to be tested. When creating learning content, design decisions are highly contextual. For instance, there is no consensus in the design research literature on whether video is useful for learners simply because the answer depends so much upon contextual factors—such as (a) the type of video, (b) the quality of video, (c) its relationship to the text, (d) the age and characteristics of the learner, etc.—and even proposing decontextualized design decisions that are intended to be universally applied (like “what are the effects of video on instruction?”) has come to be viewed as a misguided or altogether

confounded research strategy (Honebein & Reigeluth, 2021). The alternative to this is to employ research efforts in iterative, continuous improvement where a variety of strategies might be tested in deeply contextualized ways to improve learning products, such as adding or removing a specific video to a live textbook chapter. Toward this end, this study focused on six chapters in a single textbook and experimentally tested a different design change for each chapter (representing two versions of each chapter) to determine the feasibility of testing and revising these kinds of design decisions on-the-fly with live products. For instance, in the "Technology Integration" chapter, the experimental flight removed stock photos to determine whether the mere presence of photos influenced perceptions of quality. Similarly, in the "Lifelong Learning" chapter, the experimental flight removed an introductory video for the same purpose. Other changes made to remaining chapters included (a) adding extra images (for "Information Literacy"), (b) removing direct illustrative quotations (for "Online Professionalism"), (c) increasing the font size (for "Online Safety"), and (d) changing the sans-serif font style to a serif font (for "Universal Design for Learning"). In every case, chapters were set to "Automatic" flight assignment for a one-year period, and a series of Welch's t-tests were conducted to determine whether the change influenced overall quality ratings for the chapter in the target time period.

In constructing these experiments, we did not expect to see drastic differences in results, but we did anticipate that if we could identify small formatting or content changes that resulted in small quality differences, then as these changes were aggregated together and applied to the entire textbook, overall quality could be improved in meaningful ways. For instance, even if adjusting stock photos, fonts, or videos only affected less than a 10% change each in perceived quality, by applying these results to all of the chapters we hoped to be able to improve chapters in ways that would show significant aggregate benefit. Additionally, because all of these experiments reflected relatively low-cost adjustments to resources that are used by

a large number of people, even small improvements would be expected to have considerable relative advantage. For instance, if a small change can improve readability by only 1% of a textbook with a readership of 50,000, that small change could mean that 500 more people might actually benefit from the resource. Thus, though small improvements may historically be treated as insignificant in educational settings that are constantly seeking after silver-bullet or 2-sigma solutions (e.g., Bloom, 1984), when we move into the realm of high-impact open resources that we can adjust at low-cost, even tiny improvements can yield drastic results in learning for the broad population.

Results and Discussion

The simple result of this study is that after one year of constant data collection on a popular open textbook, all experiments came back as having statistically non-significant effects on perceived open textbook chapter quality. It is no secret that educational research exhibits a strong bias against reporting null effect studies, which leads many researchers to not publish valuable work and contributes to "publication bias, a positively skewed research base, and policy and practices based on incomplete data" (Cook & Therrien, 2017, p. 149), but even though results for this study were non-significant, the results may nonetheless be valuable for informing ongoing research and practice with continuous improvement efforts and open educational resources.

Table 1 provides a summary of the results for all six experiments, and there are at least two items of interest from the results that seem noteworthy. First, though non-significant, the Cohen's *d* values for several of the experiments approach levels that suggest mild to moderate strength (e.g., $d = .58$ in the case of removing the introductory video for "Lifelong Learning," and $d = .45$ in the case of switching to a serif font for "Universal Design for Learning"). Though

we cannot say for sure, these values suggest that with a larger sample size we might see effects that could mildly influence overall chapter quality perceptions, let alone aggregate effects.

Table 1

Results Summary of A/B Test Experiments for Specific Chapters

	Original Version (A)			Experimental Flight (B)						
Experiment	**Mean Rating**	**n**	**SD**	**Mean Rating**	**n**	**SD**	**Change**	**Welch's t-Test**	**p-value**	**Cohen's d**
Remove Stock Photos	4.09	256	0.7	4.19	195	0.63	0.11	1.66	NS	0.23
Remove Intro Video	4.19	70	0.66	3.95	44	0.6	-0.23	-1.92	NS	0.58
Add Extra Images	4.16	56	0.73	3.98	49	0.65	-0.18	-1.34	NS	0.38
Remove Quotations	4.26	100	0.73	4.16	88	0.6	-0.1	-1.04	NS	0.23
Increase Font Size	4.21	78	0.72	4.2	45	0.62	-0.01	-0.04	NS	0.01
Serif Font Style	4.09	58	0.7	3.88	24	0.67	-0.21	-1.29	NS	0.45

Building off of this, the second noteworthy element is the seemingly small sample size for each experiment. Though I explained this phenomenon and provided justification for why we might not expect larger sample sizes from free-roaming user behaviors above, the difficulty that this places on using these data for continuous improvement is that we seem to need an absurdly large amount of reader activity in order to collect a sufficient amount of optional self-report data for reliable testing. However, these results suggest that doing such work is feasible but that it just takes time and lots of data, especially when data are collected in unobtrusive ways and focus on user perceptions rather than discrete behaviors. Using the "Technology Integration" chapter as an example, only 1.2% of original

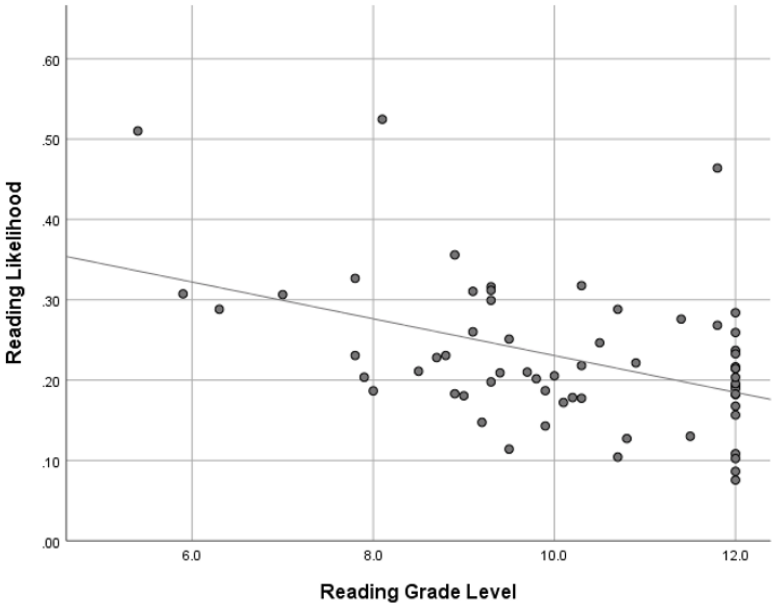
version readers and 2.0% of experimental flight readers answered the quality survey, which means that even though tens-of-thousands of users read the chapters, we still were not able to rely upon these users' data to provide sufficient evidence for improvement. This is further exacerbated by what is likely the low effect that each of these factors (on their own) has on overall perceptions of chapter quality, because smaller effects will require larger sample sizes to prove significance, and if we are only conducting experiments that we expect to have small effects, then even relatively large datasets may leave us wanting for significance. Furthermore, if these data were to be used in ongoing continuous improvement efforts, authors and researchers would find themselves in the predicament of having to throw out previous data every time they made an iterative improvement, because the original version would no longer be a valid control. The upshot of this reality is that even with a large reader base, using optional self-report data to improve open textbooks may not be a feasible approach to continuous improvement (at least not until the reader base reaches hundreds of thousands of users or more), making it difficult for most authors to make meaningful, data-driven improvements to their textbooks.

To address both of these issues, future research and development efforts would likely benefit from three key practices. First, rather than doing testing at the individual chapter or even book level, these sorts of tests might best be explored at the platform level where flights are created on all content to test for small changes. For instance, instead of removing stock photos on only the "Technology Integration" chapter, running a platform-wide flight of all chapters and programmatically removing stock photos for randomly-selected users would allow platform developers to determine the value of stock photos for EdTech Books users broadly with comparative swiftness. Similarly, doing a site-wide analysis of the effect that textual complexity has on reading likelihood reveals that likelihood goes down as complexity goes up, suggesting that as authors write chapters they should generally aim to simplify language (see Figure 5). The trade-off

with this platform-level approach is that it would lose context, because not all chapters might benefit equally from the presence or lack of stock photos due to different content and audiences and some content might require greater textual complexity, but it would at least provide platform developers with data-based guidelines to provide suggestions to authors on what effects their decisions might be having on readers (e.g., “including more than three stock photos is predicted to reduce user quality perceptions of your chapters by 11.5%”).

Figure 5

Relationship Between the Reading Grade Level of Chapters and Reading Likelihood



Picture of a Chart Showing the Relationship Between the Reading Grade Level and Reading Likelihood for Chapters

Note. R^2Linear = 0.199

Second, many of these types of tests can potentially become automated not just at the random assignment phase but also at the implementation and continuous improvement phase. For instance, if a font size experiment was implemented across an entire platform with a font-size increment of 10%, the system could create an experiment that increases font size for random users by 10% while reducing it by 10% and leaving it the same for others. This site-level test could continue until enough data were collected to determine which of the choices was optimal. In probability theory, this type of approach is called a "bandit algorithm" as it attempts to address the "multi-armed bandit problem" by maximizing positive outcomes (e.g., chapter reads, positive ratings) while simultaneously employing an exploratory mechanism to discover whether other options or features might improve results (Berry & Fristedt, 1985). Employing bandit algorithms for improving any design feature could utilize an infinite number of variables (e.g., different font sizes, types, or colors) in experimental ways that both produce actionable results and minimize undesirable outcomes. For many design decisions, this could allow continuous improvement to occur in an automated fashion without the need for authors or even developers to manually adjust designs to respond to experimental results. Rather, the design of the platform could become self-correcting in many regards to account for ongoing user behaviors.

And third, though relying on self-report data like quality ratings may still have a place (especially in larger scale analyses), more granular and faster improvements would need to rely upon unobtrusive user behavior data that is more universally collected. For instance, based on the textual complexity of a chapter and the time-on-page behaviors of a reader, we can determine whether each user actually read the page. Using this as the dependent variable would mean that we would have reliable experimental data for all learners rather than just the small subset that self-report data provides and would allow us to predict how experimental changes are affecting behaviors for all learners (e.g., does changing the font style influence the likelihood that a user will read the page?). Though this may limit our

experiments in some ways, it would allow for rapid and continuous improvement (especially when coupled with the other suggestions above) that would not be readily possible while waiting for self-report data.

Furthermore, many of these possible dependent variables would likely be correlated to one another. For instance, conducting a simple post hoc bivariate correlation of quality measures, predicted reads, and textual complexity on all chapters in the platform with at least 10 quality ratings (n = 63) revealed a significant, moderate relationship between these variables (see Table 2). This suggests that even if the primary goal is to improve perceived quality of textbooks, movement toward this goal might be accomplished in part by engaging in efforts that seek to influence more easily measurable variables (like reading likelihood).

Table 2

Bivariate Correlations of Chapter Factors

	Textual Complexity	**Reading Likelihood**
Quality Rating	**.526****	**.288***
Textual Complexity		.415**

* Denotes significance at the p < .05 level.

** Denotes significance at the p < .01 level.

Conclusion

In conclusion, though the experiments presented in this study yielded non-significant results, findings remain valuable for helping researchers and authors interested in engaging in data-driven continuous improvement efforts for several reasons. First, this study

points out the relative difficulty of engaging in these efforts at a granular level (e.g., at the chapter or resource level), especially when the resources that we are seeking to improve do not enjoy viral popularity. Rather, such efforts are likely best addressed at the system level where experimental flights may be created with, randomized for, and aggregated from many different resources at once. Second, due to the relative simplicity of many of these experimental conditions, platform developers should explore automating not just the randomization aspect of A/B tests but also the actual implementation and experimental creation of tests, allowing the system to iteratively experiment-improve-experiment in valuable directions by employing bandit algorithms. And third, because these efforts rely upon unobtrusive data collection, continuous improvement will most effectively be influenced by data that can be collected from as many users as possible without relying upon low-probability participation metrics such as prompting users to answer a survey or to provide a rating. Incorporating these suggestions into any open textbook continuous improvement effort would offer great promise for making the most of user experience data that is readily available in many open platforms today. By doing so, the theoretically achievable goal is to create continuous improvement systems that are not only comparable to traditional publishing mechanisms but that far exceed them in ensuring the usefulness, usability, and perceived quality of open resources.

References

Berry, D. A., & Fristedt, B. (1985). *Bandit problems: Sequential allocation of experiments* (Monographs on statistics and applied probability). Chapman and Hall, 5(71-87), 7-7.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16.

Bodily, R., Nyland, R., & Wiley, D. (2017). The RISE framework: Using learning analytics to automatically identify open educational resources for continuous improvement. *International Review of Research in Open and Distributed Learning, 18*(2), 103-122. https://doi.org/10.19173/irrodl.v18i2.2952

Bozkurt, A., Koutropoulos, A., Singh, L., & Honeychurch, S. (2020). On lurking: Multiple perspectives on lurking within an educational community. *The Internet and Higher Education, 44*, 100709. https://doi.org/10.1016/j.iheduc.2019.100709

College Board. (2019). Published prices. *Trends in Higher Education*. https://trends.collegeboard.org/college-pricing/figures-tables/published-prices-national

Cook, B. G., & Therrien, W. J. (2017). Null effects and publication bias in special education research. *Behavioral Disorders, 42*(4), 149-158. https://doi.org/10.1111/ldrp.12163

Dinevski, D. (2008). Open educational resources and lifelong learning. In *ITI 2008 - 30th International Conference on Information Technology Interfaces* (pp. 117–122). IEEE.

Honebein, P. C., & Reigeluth, C. M. (2021). To prove or improve, that is the question: the resurgence of comparative, confounded research between 2010 and 2019. *Educational Technology Research and Development*, 1-32. https://doi.org/10.1007/s11423-021-09988-1

Gütl, C., Rizzardini, R. H., Chang, V., & Morales, M. (2014, September). Attrition in MOOC: Lessons learned from drop-out students. In *Learning Technology for Education in Cloud, MOOC, and Big Data* (pp. 37-48). Springer, Cham.

Kimmons, R. (n.d.). About EdTech Books. *EdTech Books*. https://edtechbooks.org/about

Kimmons, R. (2015). OER quality and adaptation in K-12: Comparing teacher evaluations of copyright-restricted, open, and open/adapted textbooks. *The International Review of Research in Open and Distributed Learning, 16*(5).

Kimmons, R. (2016). Expansive openness in teacher practice. *Teachers College Record, 118*(9).

Kimmons, R., & Veletsianos, G. (2018). Public internet data mining methods in instructional design, educational technology, and online learning research. *TechTrends, 62*(5), 492-500. doi:10.1007/s11528-018-0307-4

Kohavi, R., & Longbotham, R. (2017). Online controlled experiments and A/B testing. *Encyclopedia of Machine Learning and Data Mining, 7*(8), 922-929.

Martin, M. T., & Kimmons, R. (2020). Faculty members' lived experiences with open educational resources. *Open Praxis, 12*(1), 131-144. doi:10.5944/openpraxis.12.1.987

Open Source Initiative. (n.d.). *The open source definition*. http://opensource.org/osd

Ottenbreit-Leftwich, A. & Kimmons, R. (2020). *The K-12 Educational Technology Handbook* (1st ed.). EdTech Books. https://edtechbooks.org/k12handbook

Rivard, R. (2013, March 8). Measuring the MOOC dropout rate. *Inside Higher Education*. https://www.insidehighered.com/news/2013/03/08/researchers-explore-who-taking-moocs-and-why-so-many-drop-out

Seaman, J. E., & Seaman, J. (2018). Freeing the textbook: Educational resources in U.S. higher education, 2018. *Babson Survey Research Group*.

https://www.onlinelearningsurvey.com/reports/freeingthetextbook2018.pdf

Stallman, R. (2013). *FLOSS and FOSS*. GNU Operating System. https://www.gnu.org/philosophy/floss-and-foss.html

UpGrade. (n.d.). UpGrade: An open source platform for A/B testing in education. *Carnegie Learning*. https://www.carnegielearning.com/blog/upgrade-ab-testing/

Veletsianos, G., Kimmons, R., Larsen, R., & Rogers, J. (2021). Temporal flexibility, online learning completion, and gender. *Distance Education, 42*(1). doi:10.1080/01587919.2020.1869523

Whitford, E. (July 26, 2018). Textbook trade-offs. *Inside Higher Ed*. https://www.insidehighered.com/news/2018/07/26/students-sacrifice-meals-and-trips-home-pay-textbooks

Wiley, D. (n.d.). Defining the "open" in open content and open educational resources. *iterating toward openness*. http://opencontent.org/definition/

Wiley, D., Strader, R., & Bodily, R. (2021). Continuous improvement of instructional materials. In J. K. McDonald & R. E. West (Eds.), *Design for learning: Principles, processes, and praxis*. EdTech Books. https://edtechbooks.org/id/continuous_improvement

Woodward, S., Lloyd, A., & Kimmons, R. (2017). Student voice in textbook evaluation: Comparing open and restricted textbooks. *The International Review of Research in Open and Distributed Learning, 18*(6). doi:10.19173/irrodl.v18i6.3170

Kimmons, R. (2021). A/B Testing on Open Textbooks: A Feasibility Study for Continuously Improving Open Educational Resources. In Y. Arts, H. Call, M. Cavan, T. Holmes, J. Rogers, S. Tuiloma, L. West, & R. Kimmons (Eds.), *An Introduction to Open Education*. EdTech Books. https://edtechbooks.org/open_education/ab_testing_on_open_t