Varied Practice Testing is Associated with Better Learning Outcomes in Self-Regulated Online Learning

Paulo F. Carvalho, Elizabeth A. McLaughlin, and Kenneth R. Koedinger

Human-Computer Interaction Institute, Carnegie Mellon University

Author Note

Paulo F. Carvalho https://orcid.org/0000-0002-0449-3733

Elizabeth A. McLaughlin https://orcid.org/0000-0003-2650-6504

Kenneth R. Koedinger https://orcid.org/0000-0002-5850-4768

The authors declare no competing interests. This work was supported by National Science Foundation grant ACI-1443068 toward the creation of LearnSphere.org, National Science Foundation grant BCS-1824257, and funding from Google. The authors are thankful to the editor and 3 anonymous reviewers for the helpful feedback and to the members of LearnLab for helpful discussion.

P.C. and K.K. designed and directed the project; P.C. obtained the data with support from E.M. P.C. and K.K. developed the theoretical framework. P.C. and K.K. developed the analytical approach and P.C. analyzed the data with support from E.M. All authors contributed to the writing of the manuscript.

Correspondence concerning this article should be addressed to Paulo F. Carvalho, Human-Computer Interaction Institute, Carnegie Mellon University, Newell-Simon Hall, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: pcarvalh@cs.cmu.edu

Abstract

In this paper, we leverage data from over 1,000 students participating in two different online courses to investigate whether better learning outcomes are associated with student decisions to practice instead of (re-)reading. Consistent with laboratory and classroom findings, we find that students' decisions to practice are related to better learning outcomes. Moreover, we find that this benefit is particularly related to increasing the number of different practice activities completed and not repeating the same activity multiple times. Our findings are consistent with theories suggesting that practice testing improves learning by enhancing encoding promoted by practicing the same content with different problems and raise questions regarding the benefits of repeatedly practicing the same question. The work presented here also demonstrates one way we can leverage data from naturally occurring datasets and learning analytics approaches to inform theoretical developments and the understanding of cognitive phenomena. We argue that to fully understand the cognitive processes involved in learning we need to test our hypotheses in natural educational contexts, both using controlled experimentation and analyses of naturally occurring data.

Keywords: active learning; self-regulated learning; practice testing; learning analytics; variability

Educational Impact and Implications Statement

In online and blended learning environments completing more different activities was associated with improved student learning outcomes. Contrary to previous suggestions from questionnaire evidence, these findings suggest that students in real-world environments, not just in highly controlled laboratory settings, can effectively use practice as a self-regulated learning opportunity. Best learning outcomes will be achieved from learning environments that provide ample opportunities for practice.

Varied Practice Testing is Associated with Better Learning Outcomes in Self-Regulated Online Learning

Advancements in our understanding of the cognitive processes involved in learning have led to a better understanding of how educational environments can be better structured (e.g., Bjork et al., 2013; Koedinger et al., 2012), and how learning outcomes can be improved (e.g., Deans for Impact, 2015; Dunlosky et al., 2013; Pashler et al., 2007).

Despite this progress, there are open questions about which insights garnered from laboratory-based and small randomized field trials generalize to other real-world educational problems and settings. Because real-world learning often involves the interaction of many variables that are often controlled for or eliminated in laboratory and single intervention studies (e.g., student prior knowledge, motivation, interest, preferences, learning materials, learning tasks, timescales), it is possible, perhaps likely (see e.g., Koedinger et al., 2012) that interactions between these variables lead to different results than what is found in the lab or single field trials. Furthermore, understanding whether findings from non-representative small samples generalize to real-world human activity at scale is crucial to understanding, interpreting, and acting upon such findings. One could, and perhaps should, ask whether a theory can explain human behavior if it does not encompass behaviors that happen naturally in the world.

Learning analytics of large-scale real-world learning data is an important and useful opportunity to contribute to cognitive theory with increased ecological validity and more direct relevance to practice. Data science approaches using large-scale analyses allow us to understand the boundaries and limitations of our current understanding and the generalizability of laboratory findings (e.g., Kim et al., 2019; Ridgeway et al., 2017; Stafford & Dewar, 2014; Steyvers & Benjamin, 2019). For example, recent research using big data has expanded our understanding of the spacing effect found in the laboratory (e.g., Carvalho et al., 2020; Kim et al., 2019; Stafford & Dewar, 2014). Stafford and Dewar (2014) used a large sample of people playing an online game to demonstrate that spaced practice is related to better performance at scale. Similarly, Kim and collaborators (Kim et al., 2019) used data from workplace training materials to demonstrate that, as suggested by laboratory studies, the optimal interval between repetitions increased as the retention interval increased. Recently, Carvalho et al. (2020) used data from a Massive Online Course to demonstrate a self-regulated spacing effect: when students spaced their activities across time they performed better in the exams compared to when the same students spaced their activity less. These examples highlight how using large datasets that come from multiple real-world settings increases the external validity of our theories, models, and conclusions in combination with laboratory evidence.

Practice Testing

When students practice by completing tests as opposed to reading (or generally any other “studying” behavior using other approaches such as concept mapping, reading, re-reading, see e.g., Karpicke and Blunt, 2011), they show better learning outcomes, particularly in a delayed test, a phenomenon known as the testing effect, retrieval practice, or practice testing. There are multiple theoretical explanations of the testing effect. Extant theories are described in abstract terms and often focus on only a subset of phenomena specific to the experimental manipulation at hand. Therefore, existing theories are not necessarily mutually exclusive or complete. A subset of theories that has received some support from meta-analyses of existing laboratory data (Rowland, 2014) can be referred to as retrieval effort theories (e.g., Pyc & Rawson, 2009). The general proposal of these theories is that the effect arises from increased effort or depth of processing during practice (e.g., Bjork & Bjork, 1992; Pyc & Rawson, 2009), although the exact mechanism whereby this takes place varies. Thus, one of the main predictions of this group of theories is that increasing the number of practice tests should yield larger practice benefits (Glover, 1989). Other theories suggest that the benefits of retrieval practice relate to the retrieval of the same information that takes place with additional testing (Benjamin & Tullis, 2010; Carpenter, 2009). According to these theories, each repeated attempt to retrieve the *same* information changes the memory trace by, for example, increasing the routes to retrieval or adding additional context.

Current evidence of the benefits of practice testing comes in large part from small scale laboratory studies. In fact, only 15% of the studies reviewed in a recent meta-analysis were conducted in a classroom setting (Adesope et al., 2017; see also Pan & Rickard, 2018). Furthermore, a recent classroom study found a reverse effect—better performance for students who completed fewer practice tests (Gurung & Burns, 2019)—consistent with prior theoretical analyses in the context of problem-solving activities instead of retrieval of information (Koedinger & Aleven, 2007). The absence of a practice testing effect has also been reported when using self-explanation as a control condition instead of the typical reading/re-reading control condition in laboratory studies of the testing effect (Aleven & Koedinger, 2002). Moreover, laboratory experiments question the effectiveness of practice testing when the question practiced is not repeated in the final test (Nguyen & McDaniel, 2015; Wooldridge et al., 2014).

Thus, real-world evidence for the benefits of practice testing is, at best, mixed, potentially lacking. Moreover, although there has been some initial investigation of how students in natural contexts self-regulate practice between study and testing (Andergassen et al., 2014; Kizilcec et al., 2017), these inquiries have been limited in scope both in terms of duration of practice (e.g., only 14 days) and type of materials (e.g., only multiple-choice testing questions). Given past results and theoretical analyses suggesting such variables are important to whether practice testing improves learning or not (e.g., Rowland, 2014), it is an open question how far practice testing extends to real-world situations where students' previous knowledge and characteristics, types of materials, and perceived importance vary substantially, unlike in the lab.

Moreover, evidence that practice testing—in the laboratory or outside of it—improves learning when students *decide* to engage in it through self-directed practice is even scarcer. Most studies investigating the benefits of practice testing do so in the context of *directed instruction*, that is, situations where demand characteristics produce essentially full compliance in testing or reading/studying, and often without accounting for instructional time (for reviews see Roediger & Butler, 2011; Roediger & Karpicke, 2006a). Although there is evidence that telling students about the benefits of retrieval practice can improve self-regulated decisions of when to read, practice or stop altogether (Ariel & Karpicke, 2018; Yang et al., 2017) and some evidence that students might not realize the benefits of practice testing when self-regulating their study (Janes et al., 2018; Son & Kornell, 2009; Toppino et al., 2018; Wissman et al., 2012), there is currently no clear evidence that students benefit from practice testing when they spontaneously can choose to take such opportunities for themselves. This gap is critical to our understanding of the phenomenon because self-regulated learning often changes more than just who is in control. From the perspective of the learner, unlike directed learning, self-regulated learning is not based on randomly sampling information from a distribution (Castro et al., 2009; Gureckis & Markant, 2012). When learners regulate their learning well, they can target their practice to areas of uncertainty, which allows for capture of optimal novelty and prediction error reduction (that is, better connecting what they know with what they do not yet know, so as to improve ability to make correct predictions/responses in the future; Gureckis & Markant, 2012). In fact, students' decisions to re-read/study instead of practice testing can yield better learning than forcing them to practice (Carvalho et al., 2018; Tullis et al., 2018). Furthermore, it has been highlighted before that best-practices identified in directed learning do not always match best-practices for self-regulated learning (Carvalho et al., 2016; Ciccone & Brelsford, 1976). Thus, there is a need to better understand whether self-regulated testing practice influences learning outcomes in the same ways as directed testing practice does. It is important to note here that although self-regulated learning is multifaceted and at times hard to define (e.g., Chi, 2009), here we focus only on one dimension of self-regulated learning: the learning consequences of allowing students to decide whether to practice or not.

Variability of activities in practice testing

The effect of practice testing on learning outcomes has classically been studied by comparing practicing the same activity multiple times with rereading/studying the same information the same number of times (see e.g., Roediger & Karpicke, 2006b; Rowland, 2014). Thus, it is currently an open question whether practice testing benefits learning by improving memory for specific questions and responses or whether practice testing can benefit learning by improving memory for common knowledge across different questions.

Although, there is some evidence from laboratory work suggesting that, compared to re-studying, practice testing without repetition can lead to improved learning (Agarwal, 2019; Jensen et al., 2014), and that practice testing with repeated questions can improve not only memory but also generalization across test formats (Butler, 2010; Kang et al., 2007), to date there have been no direct evaluations of the effectiveness of practice testing when the repeated practice tests vary. This question is particularly important in applied contexts. Educators, students, and educational technology developers might wonder whether to promote learning and retention with practice testing, repeated practice with the same question is needed, or varied practice is sufficient or even preferable.

As mentioned above, some theories suggest that the benefits of retrieval practice are connected with the retrieval of the same information that takes place with additional testing (Benjamin & Tullis, 2010; Carpenter, 2009), whereas others suggest that the benefits are more connected with how information is initially encoded during practice testing (especially after receiving feedback) compared to reading situations (Pavlik & Anderson, 2008; Pyc & Rawson, 2009). If the benefits of practice testing are connected with retrieval of the same information, we would expect that repeating the same practice test multiple times would result in particularly high posttest performance because it emphasizes repeated *retrieval*, which would lead to better future recall and generalization (Wissman et al., 2018). In contrast, if the benefit of practice testing relates to differential encoding of the information compared to reading, we would expect better posttest performance to relate to variety in testing opportunities that target the same content in different ways allowing for *elaboration and encoding* of the critical information and inhibition of the non-critical information. There is a wealth of research on generalization and extrapolation suggesting that variability during study improves learning (e.g., Lively et al., 1993; Posner & Keele, 1968). One possible explanation is that studying multiple examples with varying surface characteristics, but similar underlying structure, promotes transfer because it allows learners to identify relevant properties (the underlying structure) and ignore irrelevant surface variation. That is, the underlying solution can be abstracted as a generalized solution or schema (e.g., Catrambone & Holyoak, 1989; Gick & Holyoak, 1980, 1983; Holyoak & Koh, 1987). For example, Paas and VanMerrienboer (1994) found that when learning geometry problems through worked-out examples, learners' transfer of their learning to novel problems at test was better following study of many varied examples compared to less varied examples. Interestingly, in that study, more variable examples were associated with learners devoting more time studying the problems and potentially encoding the common solution better. It is unclear whether such benefits generalize to self-regulated practice testing situations or whether the testing effect is tied to repeatedly working on the same activity.

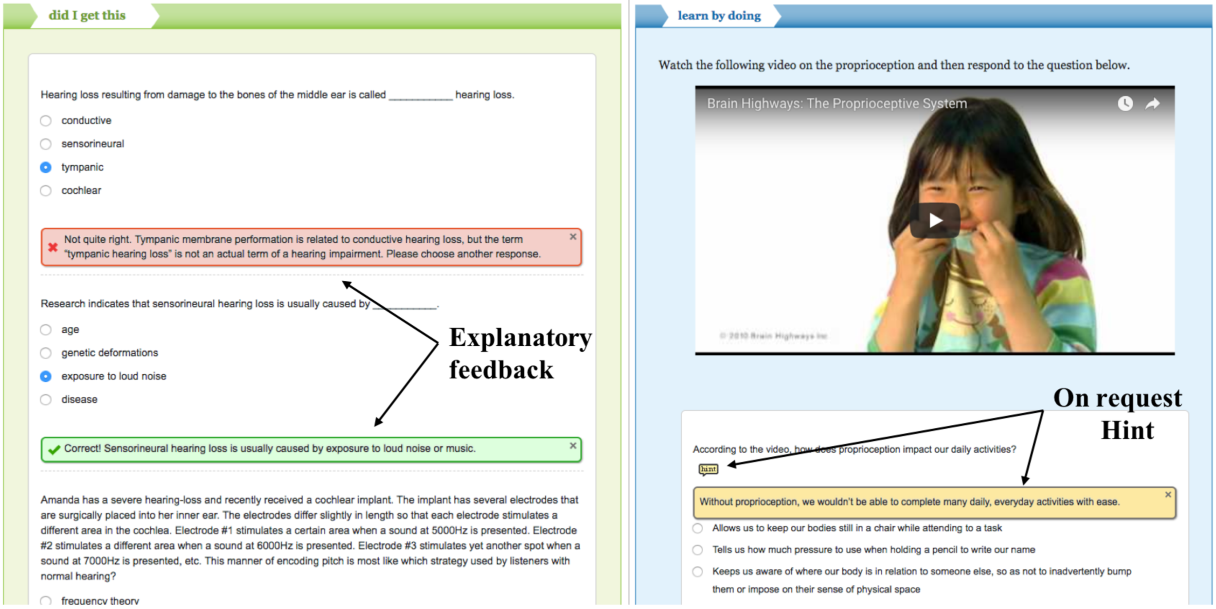
The Open Learning Initiative environment

The datasets we used in the current studies come from courses using materials developed by the Open Learning Initiative (OLI) at Carnegie Mellon University. OLI is designed to support better learning and instruction by offering open and free courses that contribute to research efforts for improving course development and design through an iterative process using real student data. Through backward design approaches, OLI courses are developed such that each activity, page, and assessment presented to students is aligned with a learning objective. This process results in highly integrated materials and the system tracks students' progress on each learning objective (Bier et al., 2011). This approach has been shown to result in highly effective courses that promote learning (Lovett et al., 2008).

The structure of OLI courses, like a textbook hierarchy, modules, learning objectives, and practice activities. Each webpage of an OLI course might include explanatory text and/or practice activities. In fact, multiple resources are usually shown on the same page. The fundamental and important difference between OLI and textbook learning is that OLI provides (1) an interactive environment with immediate explanatory feedback during practice activities and (2) formative assessments embedded in the course as students are learning. OLI modules include a variety of expository content (text, examples, images, and video clips) and many interactive activities, all addressing common learning objectives. Broadly, these activities serve two purposes. "Learn By Doing" activities, intended to support student outcome achievement, provide feedback and detailed hints. Another type of activity, "Did I Get This," provides a self-comprehension check for students with feedback but no hints. The interactive activities were created in conjunction with the OLI text materials and complement it by providing testing ("Did I Get This") or active learning ("Learn by Doing") activities that cover the same concepts described in the text. Figure 1 presents an example of each of these types of activities and their properties, including hints and feedback.

Figure 1

Examples of activities in the OLI online textbooks.



Hearing loss resulting from damage to the bones of the middle ear is called __________ hearing loss.

○ conductive
○ sensorineural
● tympanic
○ cochlear

✖ Not quite right. Tympanic membrane performation is related to conductive hearing loss, but the term "tympanic hearing loss" is not an actual term of a hearing impairment. Please choose another response. ×

Research indicates that sensorineural hearing loss is usually caused by __________.

○ age
○ genetic deformations
● exposure to loud noise
○ disease

✓ Correct! Sensorineural hearing loss is usually caused by exposure to loud noise or music. ×

Amanda has a severe hearing-loss and recently received a cochlear implant. The implant has several electrodes that are surgically placed into her inner ear. The electrodes differ slightly in length so that each electrode stimulates a different area in the cochlea. Electrode #1 stimulates a certain area when a sound at 5000Hz is presented. Electrode #2 stimulates a different area when a sound at 6000Hz is presented. Electrode #3 stimulates yet another spot when a sound at 7000Hz is presented, etc. This manner of encoding pitch is most like which strategy used by listeners with normal hearing?

○ frequency theory

Watch the following video on the proprioception and then respond to the question below.

According to the video, how does proprioception impact our daily activities?

Without proprioception, we wouldn't be able to complete many daily, everyday activities with ease. ×

○ Allows us to keep our bodies still in a chair while attending to a task
○ Tells us how much pressure to use when holding a pencil to write our name
○ Keeps us aware of where our body is in relation to someone else, so as not to inadvertently bump them or impose on their sense of physical space

Note. Screenshot of an OLI "Did I Get This?" activity (left panel) and a "Learn by Doing" activity (right panel) in the Psychology course. Detailed explanatory feedback is immediately available once the student clicks or submits an answer, both for correct and incorrect responses (see left panel). Hints are available in "Learn by Doing" activities by clicking the "Hint" icon (see right panel). See also Supplementary Materials for further examples.

A team of experts and learning engineers developed the OLI materials and tagged every activity and page with the learning objectives it targets, allowing instructors to track students' performance and progress on each learning objective. For example, the learning objective "Identify the structure and function of the cerebellum" in the Psychology course is associated with the pages where text related to that objective is covered as well as any activity that assesses that objective. Each learning objective can be associated with one or more pages and one or more activities. As students complete each activity, their performance informs a student learning model that tracks progress towards mastery of each learning objective. Instructors have access to this information through the teacher dashboard.

Furthermore, this tight integration and focus on alignable learning objectives allows us to compare students' outcomes from engaging with activities and reading/studying *the same content*. Because pages and activities are tagged with the same learning objectives, we have information about the content covered and whether students engaged with reading or activity materials for the same content, which is not always possible with naturalistic data, hindering investigations of self-regulated practice testing effects in natural contexts.

The DataShop repository

The datasets used are available through DataShop, an open learning repository for educational data (Koedinger et al., 2010). DataShop is a central repository for educational datasets specializing in data from the interaction between students and educational software, including data from online courses, intelligent tutoring systems, virtual labs, online assessment systems, collaborative learning environments, and simulations. The data available through DataShop are fine-grained (click by click), longitudinal (entire semesters or years), and extensive (multiple datasets for the same software or online course provider), allowing the type of detailed student behavior analysis presented here.

The current study

In this paper, we investigated the existence and characteristics of self-regulated practice testing. We did so by analyzing data available in a repository of educational data (DataShop). We defined practice testing as completing optional activities in a course. First, we established whether completing more activities relates to best learning outcomes. Then, we analyzed the characteristics of this relationship. We analyzed two dimensions of variability: (1) whether best learning outcomes were associated with practicing more learning objectives (increased coverage) or repeatedly practicing the same learning objective (drill practice); and

(2) best learning outcomes were associated with practice with different questions or repeatedly completing the same question. As mentioned above, taken together, these analyses allow us to identify whether the benefits of self-directed practice are related to increased effort or retrieval

Study 1: Psychology Course

Data and Methods

Transparency and Openness

We report how we determined our sample size, all data exclusions, and all measures in the study, and we follow JARS (Kazak, 2018). All data is available at Datashop.org (https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=863) and all analysis code is available at GitHub ([removed to comply with masked review policy]). Data were analyzed using R, version 4.1.1 (R Core Team, 2021) and the following packages: lme4, version v1.1-14 (Bates et al., 2014), ggplot2, version 3.3.2 (Wickham, 2016), and EMAtools, version 0.1.3 (Kleiman, 2021). This study's design and its analysis were not pre-registered.

Ethics

Data collection for this Study and Study 2 was approved under Carnegie Mellon University Institutional Review board (CMU IRB) protocol #HS11-351. The DataShop repository and its use is approved under CMU IRB protocol #IRBSTUDY2015_00000236. As the data are archival and anonymous, there was no written informed consent required. Written consent was waived because participation in the course was part of normal educational practices. As per DataShop requirements, all available data were verified for appropriate student participation agreement and IRB oversight, and students provided consent to have their data analyzed.

Course and participants

In this study, we analyzed data from an online course “Introduction to Psychology as a Science” taught through Coursera. The course was open to the public. It was a 12-week course, taking place in 2013. Students received credit for completion if they completed quizzes and the final exam. A total of 5,615 students enrolled in the course and agreed to participate. Of these, 851 completed the final exam, 3,914 completed the pretest, and 829 completed both exam and pretest data. Completing the exam and most of the quizzes was a requirement to complete the course, thus, we considered data from only students who completed the final exam and the pretest (N = 829, a 15% conclusion rate, in line with other MOOCs, see e.g., Liyanagunawardena et al., 2014). Of these 829 students, 74 did not have any interaction data in the system and were thus excluded from analyses. No other inclusion criteria were used. The final sample included in analyses included 755 students who completed the course. Of note, students who did not complete the exam were also unlikely to complete the pretest, often only completed only one of the quizzes and did not interact with the materials (so measures of reading and activity completion could not be derived). All students included in the analyses completed at least one of the quizzes, and most students in our sample completed all 11 quizzes (*N* = 640). We focused on students’ use of the OLI textbook materials and how it relates to performance in the quizzes. Data for this course were retrieved from DataShop (Dataset 863). For details on the number of students in the full sample and those included see Table S1 in Supplementary materials).

Students were expected to watch video lectures and complete the related textbook materials developed by OLI. OLI’s Introduction to Psychology course helps students learn key issues and theories that underlie the domain (e.g., personality, perception, research methods, development). The course is equivalent to a one semester college introductory course structured to cover learning objectives across 11 modules.

Assessments

Students were evaluated using two written assignments (40% of their grade), 11 quizzes (30% of their grade), and a final examination (30% of their grade). Students were able to take each quiz only once. An overall grade of 70% or above was considered a passing grade.

Each week, on Friday morning a multiple-choice quiz was made available. This quiz tested students on the content of that weeks' materials (thus, a quiz per module except the last module, which was covered only in the final exam). Quizzes were not timed, and students had a week to complete the quiz (but only one try). Multiple students were granted extensions to complete the quizzes. A pretest was given at the beginning of the course and a quiz was taken at the end of each module. The pretest included a series of true/false questions about the students' general knowledge of psychology and each quiz included a series of multiple-choice questions about the modules' content.

The final exam covered material from the entire course and took place in the last week of the course. The final exam was available for 5 days and students were able to take it only once.

Measures of student resource use

From the logs available we extracted, for each student and module, a series of measures related to their resource use (i.e., activity completion and page viewing). In addition, we extracted other contextual variables (activity performance, repeated vs. unique activities), information on the learning objective each activity targeted, as well as pretest and outcome measures (quizzes).

The OLI platform provides a unique identifier for each activity and page in a course and course transaction logs indicate when an activity is started and when a page is clicked on. Thus, every time a student clicked a page a log was created and every time a student interacted with an activity a log was created. The count of each of these logs is a straightforward way to operationalize page and activity use, thus we computed frequency counts for each activity completed and each page accessed. However, because activities were embedded in the pages, it is possible that some of the page accesses do not involve any reading and the student accesses it only to complete the activities. To minimize this possibility, we eliminated from page view counts any page accessed for a short period of time (bottom 5% of the distribution of the difference between page view start and activity start for all students), to reduce the possibility of counting as page viewing an event in which students did not view the text but instead moved directly to the activity. This threshold was defined based on initial inspection of the distributions of total time on each page to identify typical reading time. Moreover, if the timestamps for activity completion and page change matched, we did not count that page as a reading activity, but if they did not match (that is, if time elapsed between completing the activity and moving pages or logging off) then we assume that some reading took place and count that page as reading as well. Of course, it is possible that a student opened a page, stayed there for a while, and then completed an activity without reading the text. Although we do not have a way to account or control for this possibility, it is an equally likely possibility in most studies involving reading, even in the laboratory (e.g., Rothkopth,1968): the presentation of text presupposes reading, even if reading does not take place. It is important to note that doing activities was not a requirement in either of the courses and students were allowed to repeat the same activity as many times as desired.

Using the unique identifiers of course content, we calculated the total number of pages and activities available in each module. The total number of available pages and activities for each module was then used to calculate the percentage of activities/pages each student accessed (see Table 1 for the number of available resources per module). Because a student might complete all activities/pages and repeat an activity/page multiple times, this percentage might be larger than 100%. Although repetition is also possible even if the percentage is lower than 100%, when a student repeats only a few activities and does not complete all available activities. For this reason, when comparing activities to pages across modules we use percentage accessed from the total available. However, when investigating the amount of repeated completion of the same activity we use raw count values instead of percentages, as repetition is unlikely to be affected by the total number of available activities.

For each module, we also calculated how many activities and pages students accessed in other modules by summing up activities and page accesses on previous and subsequent modules. For example, for module three, we added up all activities accessed by that student for modules 1-2 and 4-11 and divided by the sum of activities available across modules 1-2 and 4-11. This percentage of activities/pages completed outside of the target module (module three in the example) will be used to predict performance in the target module (see below for details).

Learning Objective tagging

Learning objectives were mapped by the OLI developers to each activity and page available to the students. Logs indicated which learning objectives the activities were tagged with.

Data Analyses

Analytic approach. We used linear mixed-effects regression models to analyze the data. Unless otherwise stated, all models included pretest and performance in the activities as predictors, as well as crossed random effects for student and module. These random effects specify student and module specific intercepts, under the assumption that different modules vary in difficulty and different students differ in overall ability and resource use. These random effects are included because they create nonindependence in the data, as we will analyze the relation between resource use and quiz performance for each student (repeated across modules) and module (repeated across students). For all models, we checked the assumptions of linear regression (Linear relationship, Independence, Homoscedasticity, and Normality) and no corrections were necessary.

To determine the best-fitting models, we used chi-square tests comparing models with and without the factor of interest. For interactions, we compare equivalent models with and without the interaction. For all analyses, we report standardized coefficients and confidence intervals from the full model (the model with all predictors analyzed) and model comparison chi-square tests comparing models with and without the factor of interest. All models are listed in Supplementary Materials and referenced by number in the main text. Model numbers are sequential, with later letters representing models that include all predictors in previous models plus the critical predictor added in that model (for example, model m1c includes all predictors that m1b included plus a critical predictor of interest stated in the text). To standardize measures across students and modules, we transformed all variables into z-scores. Z-scores were calculated using the mean and standard deviation for each measure for all students included in the sample. This approach also allows us to interpret regression estimates as effect sizes and compare the relative effect of different predictors. Moreover, percentages were used for activities and pages completed instead of raw numbers of activities/pages completed to account for variation in available resources across modules.

Outcome measures. Our outcome measure in this Study was performance on each of the module quizzes. Because the quizzes took place through Coursera and not OLI, we did not have access to question-by-question performance on the quizzes or learning objective tagging for the quizzes; therefore, we used average performance in each quiz as the outcome measure in this study.

We used pretest scores as an overall measure of student readiness in the course and to control for potential general ability factors. We used performance in the activities as a measure of the difficulty of the module to control for the possibility that students' behavior varied systematically with module difficulty under the assumption that harder modules involve lower performance in the activities.[1]

Results and discussion

Overall resource use

Table 1 includes descriptive statistics for resource use. Although a smaller number of students completed repeated activities, there were students repeating activities across all modules.

[1] Some students, for some modules, did not complete any course activities. For those students, activity performance was marked as 0 for the purpose of these analyses. Removing the student from the analyses did not change any of the results presented or the conclusions.

Table 1

Descriptive Counts of Available Resources, Number of Students Who Completed At least One Resource, Average Number of Resources Completed, and Percentage of Resources Completed for Study 1.

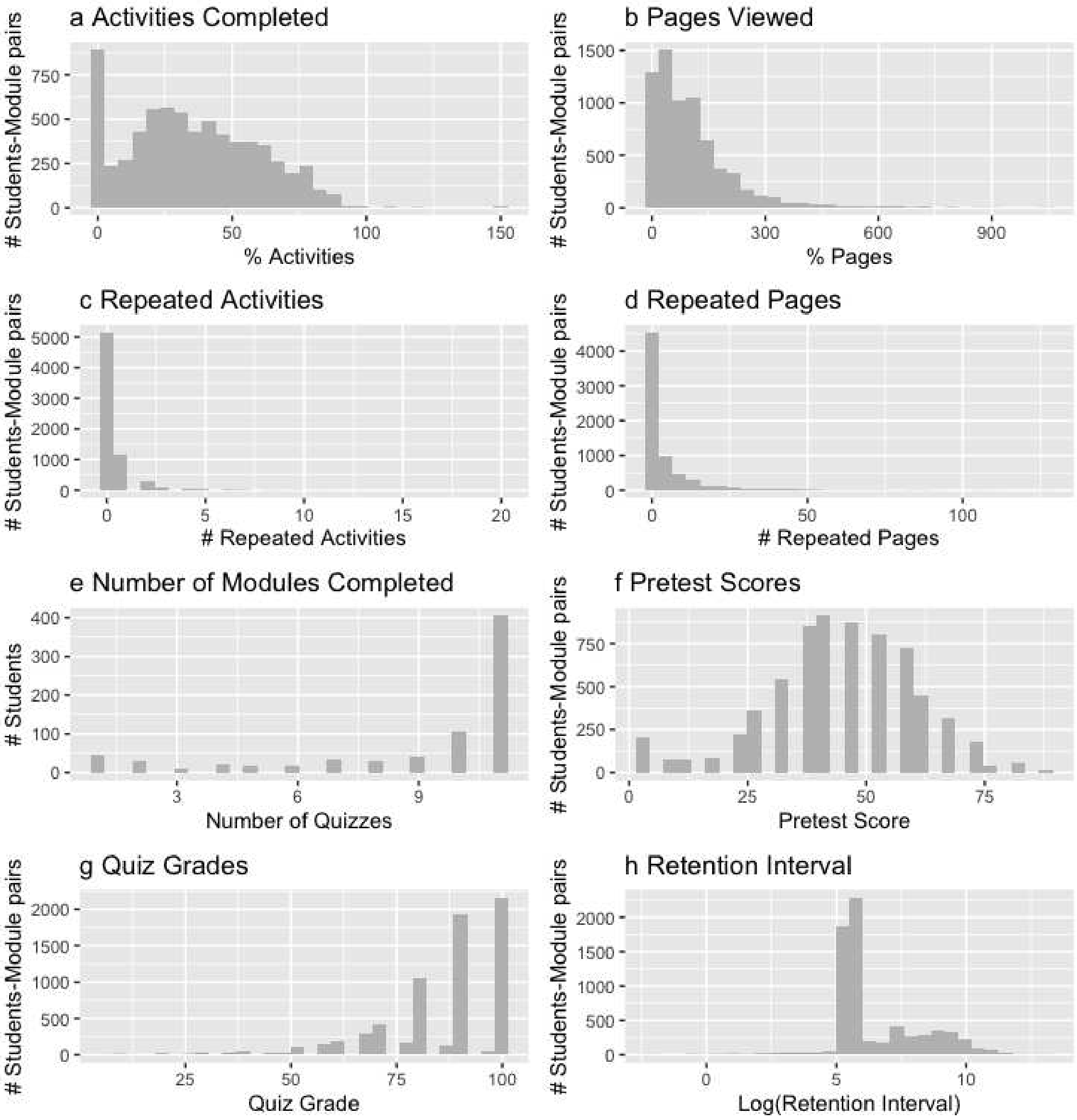
Module	Available Resources			% Students who completed at least one			Average number completed			Average % Completed		
	Activities	Pages	Learning Objectives	Activities	Repeated Activities	Pages	Activities	Repeated Activities	Pages	Activities	Pages	Learning Objectives
1	89	24	17	92%	46%	88%	41.94	0.85	12.40	47%	52%	83%
2	74	17	30	94%	35%	75%	46.96	0.54	10.60	63%	62%	84%
3	76	33	27	92%	27%	92%	33.80	0.40	19.65	44%	60%	81%
4	63	12	13	94%	30%	68%	45.19	0.43	6.05	72%	50%	87%
5	29	15	21	90%	15%	88%	14.09	0.20	10.59	49%	71%	80%
6	75	21	20	89%	27%	81%	44.24	0.40	10.29	59%	49%	76%
7	74	31	30	89%	24%	88%	36.03	0.32	15.53	49%	50%	75%
8	39	17	20	87%	14%	78%	18.82	0.17	9.13	48%	54%	76%
9	35	11	8	83%	15%	83%	20.70	0.17	7.72	59%	70%	75%
10	56	13	14	85%	25%	77%	31.66	0.35	7.03	57%	54%	75%
11	9	5	5	78%	4%	63%	5.17	0.04	2.04	57%	41%	78%

Note. N = 755. Available resources: counts of each type of resource available in the course. % students who completed at least one: percent of students (out of the total number of students available) who completed at least one of the resources available for that module. Average number completed: average number of resources completed for each module across all students; Average % completed: average (across students) percentage of available resources that were completed by students.

Initial inspection of the data (see Table 1 for summary) indicated that students completed on average 55% (SEM = 0.37%, range = [0%,111%]) of the activities and 56% (SEM = 0.79%, range = [0,680%]) of the pages available across all modules. Students completed on average 31 activities per module only once (*SEM* = 0.27, range = [0, 74]), with an average of 0.36repeated activities per module (SEM = 0.01, range = [0, 20]). Conversely, students studied an average of six pages only once (SEM = 0.07, range = [0, 28]), with an average of four repeated pages (*SEM* = 0.11, range = [0, 127]). Mean performance was 44% on the pretest quiz (SE = 0.2%, range = [2%, 87%]) and 85% on module quizzes (SE = 0.1%, range = [6%, 100%]). On average, students waited one hour and 12 minutes between finishing working on the textbook and taking the quiz (*SEM* = 2 minutes, range = [0, 110 hours]). Visual inspection of these measures, as indicated in Figure 2, shows good distribution and variability, allowing for the type of individual differences analyses we planned

Figure 2

Histograms showing distribution and variation of the different measures for the psychology course (Study 1)



Note. Counts of student-module pairs (i.e., number of data points) are presented for each predictor measure used in the regressions (activities completed, pages viewed, repeated activities, repeated pages, pretest score), and outcome measure used in the regression models (quiz grades). Number of quizzes completed by students is presented in panel e and retention interval (log transformed) is presented in panel h.

Moreover, the number of activities completed, and the number of pages viewed are not representative of stable student characteristics, rather they varied across modules for the same student. Besides visually inspecting distributions for all students, we calculated a standard deviation for the number of activities completed and pages viewed for each student. The standard deviation of the number of activities completed across modules for the same student ranges between zero and 43, with a mean of 17.39. In other words, on average, there is a wide variation in the number of activities completed across modules for the same student. This variation is higher than what is observed across students for the same module; the standard deviation of number of activities completed per module across participants ranges between three and 25, with a mean of 16. The pattern is similar for the number of pages viewed: ranging between zero and 54, with mean of 8 across modules for the same student but ranging between four and 27 across students for each module. Moreover, variation was similar for activity and read behaviors; the average coefficient of variation (a common measure of dispersion of data points, obtained by dividing the standard deviation by the mean) for pages read was 100% and for activities completed it was 78%. Given this variability, all our analyses compare performance of the same student across modules.

By considering repeated measures of different use for the same student instead of a single use-outcome score per student through the inclusion of random effects for student and module, we hope to reduce the potential impact of individual students' stable characteristics such as overall student ability or motivation. To further address this issue, we include an analysis of the percentage of resources used in other modules to predict the targeted module's accuracy. If activity/page completion is a stable student characteristic, then both the percentage of resources accessed in a target module (e.g., module three) and the percentage of resources accessed in all other modules (modules 1-2, and 4-11) will be equally good predictors of performance in the target module (e.g., module three). If, on the other hand, outcomes of a target module are related with student decisions in the target module, then the percentage of resources accessed in a target module will be a better predictor of performance in the target module than the percentage of resources accessed in other modules.

Comparing how many activities students completed in a module with how many activities students completed in all other modules, further confirms variability in student activity usage. As shown in Table 2, it is not the case that activity completion was a stable characteristic of students (the diagonal line), and many students who completed a lot of activities in other modules completed few in a particular module and vice-versa.

Table 2

Frequency of student-module pairs by activity quartile for activities completed within a module or for all other modules (outside module) in the Psychology Dataset (Study 1).

Module Usage Quartile	Outside of Module Usage Quartile				Module Totals
	25-50	50-75	Bottom 25%	Top 25%	
25-50	334	640	448	274	1,696
50-75	79	422	634	559	1,694
Bottom 25%	1,248	332	77	44	1,701
Top 25%	40	302	535	816	1,693
Outside Module Totals	1,701	1,696	1,694	1,693	6,784

Note. N = 755. Number of students-modules pairs (data points) in each quartile of activity usage for module activities (rows) and activity usage on other modules (columns). Data points in the diagonal line suggest consistent activity usage, that is, similar activity usage in the target module and other modules. Most students vary in how many activities they complete from module to module, as suggested by the variation in frequency away from the diagonal line.

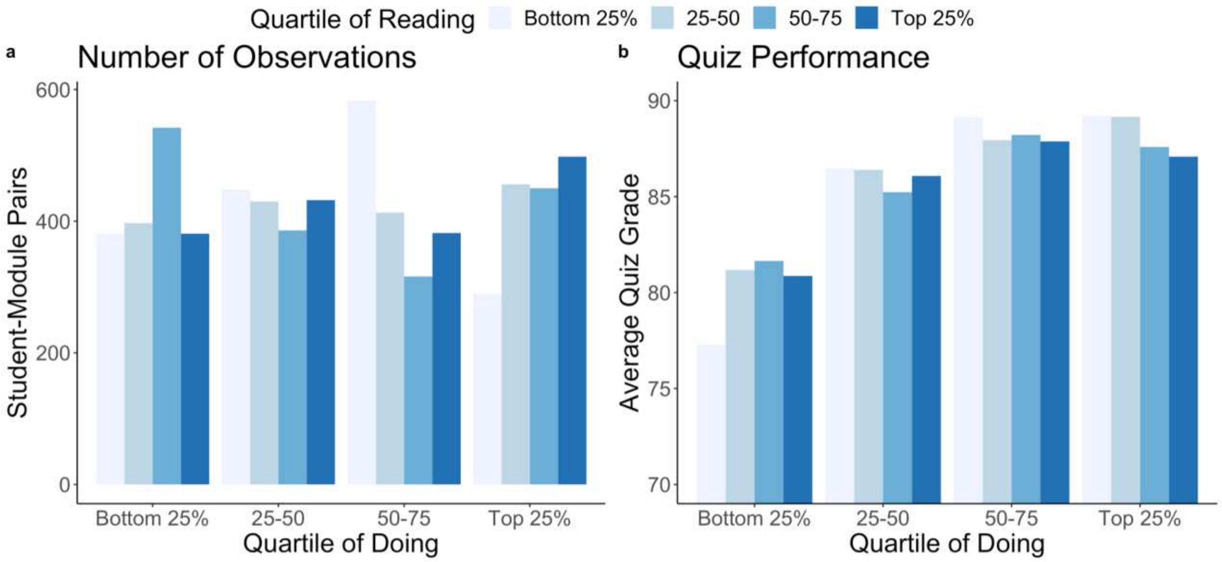
Dosage of practice and learning outcomes

We investigated the relationship between the number of practice opportunities a student engages with and their learning outcomes. Better performance in the module quizzes was positively related to better performance in the activities (m1b), $\beta = 0.07, CI: [0.039,0.093], \chi^2 = 137.95, p < .0001, d = 0.12$. Controlling for pretest scores and performance in the activities, better module quiz grades were positively related to completing more activities (m1c), $\beta = 0.16, CI: [0.131,0.195], \chi^2 = 132.46, p < .0001, d = 0.25$ (see Fig 3). Comparatively, generally completing more activities on other modules in the course had a weak relation to performance in the target module's quiz (m1d), $\beta = 0.06, CI: [0.019,0.100], \chi^2 = 5.56, p = .010, d = 0.15$. Conversely, completing more pages was not related to quiz performance (m1e), $\beta = 0.03, CI: [0.002,0.052], \chi^2 = 2.84, p = .091, d = 0.05$ (see Fig 3), whereas completing more pages on other modules was negatively related to performance in the target quiz (m1f) $\beta = -0.04, CI: [-0.087, -0.003], \chi^2 = 5.09, p = .024, d = -0.13$. The interaction between completing more activities and more pages on module quiz grades was not statistically detectable (m1e) $\beta = -0.02, CI: [-0.043,0.002], \chi^2 = 3.25, p = .071, d = -0.04$ (see Fig 3).

Thus, completing more activities is related to better learning outcomes whereas completing more pages is not. Moreover, the relation between completing more activities on the other modules was weaker or not present, suggesting that these results are not because of an overall preference of higher performing students to complete more activities on all modules.

Figure 3

(a) Number of students-module pairs with different amounts of doing and reading and (b) quiz performance as a function of amount of doing and reading for the psychology course (Study 1)



Note. The number of cases in each group in panel (a) along with the raw data for quiz grade are presented in panel (b) so the amount of data for each group can be considered. In addition, although the raw quiz grades are presented, other factors not represented in the plots (i.e., pretest scores) were considered in the regression analyses (see text for details).

Relation between learning objective coverage and quiz scores

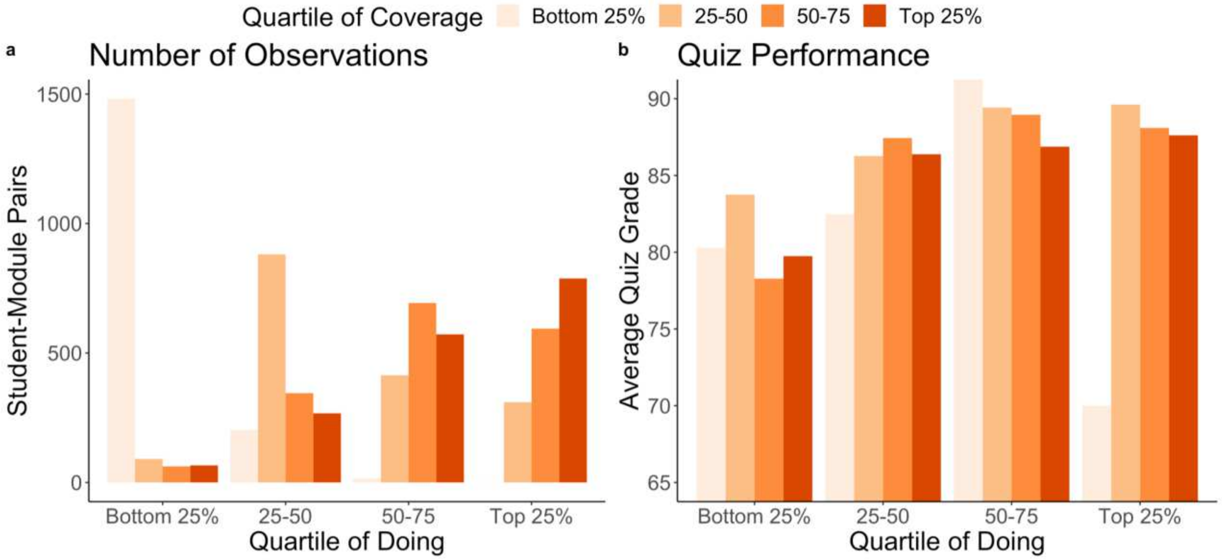
Completing more practice activities might improve learning outcomes by either increasing practice coverage (how many learning objectives were practiced), or by increasing practice drilling (how many practice opportunities there were for each learning objective). To investigate these two possibilities, we used the learning objective tagging for each OLI activity to determine what content it focused on. We calculated for each student the percentage of learning objectives they practiced (out of all the learning objectives available in each module).

We then repeated the same regression models as in the previous analyses but added the number of learning objectives practiced as a predictor.

We found that the proportion of learning objectives covered by the activities completed did not have a statistically detectable relation to quiz grades (m2b), $\beta = 0.124, CI: [0.039,0.209], \chi^2 = 0.360, p = .549, d = 0.12$, suggesting that coverage of learning objectives that would be assessed in the quiz by itself does not seem particularly related to quiz scores. There was a statistically detectable interaction such that completing more activities was more strongly associated with better quiz grades when the activities covered more learning objectives (m2c; see Figure 4), $\beta = 0.075, CI: [0.026,0.124], \chi^2 = 8.96, p = .003, d = 0.07$. Thus, completing more activities from as many learning objectives as possible was particularly related to better quiz outcomes.

Figure 4

(a) Number of students-module pairs with different amounts of doing and coverage and (b) quiz performance as a function of amount of doing and coverage for the psychology course (Study 1).



Note. The number of cases in each group in panel (a) along with the raw data for quiz grade are presented in panel (b) so the amount of data for each group can be considered. In addition, although the raw quiz grades are presented, other factors not represented in the plots (i.e., pretest scores) were considered in the regression analyses (see text for details).

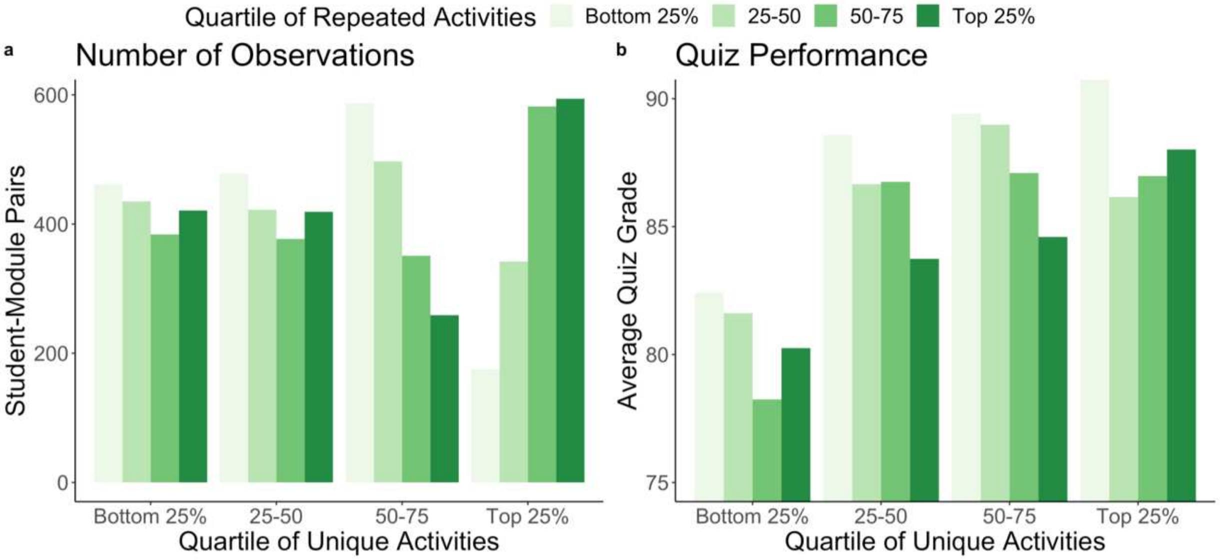
Relation between completing more different or the same activity more times and quiz outcomes

To further explore the dosage finding identified in the previous analyses, we focused on the types of practice students engaged. One possible mechanism by which practice might improve learning outcomes is by increasing effortful learning opportunities. Thus, if increased learning difficulty is the mechanism, then not only would practice be associated with better learning outcomes than more passive and less effortful approaches such as reading (as seen in the analyses above), but the effects of practice should be particularly pronounced when more effortful activities are completed compared to less-effortful opportunities. To investigate this possibility, we compared the relation between repeated practice with the *same* activity, increased practice with *different* activities, and quiz outcomes. Our reasoning is that completing the same activity and a different activity differ in the effort required. Feedback is presented after every activity and thus completing the same activity a second time would be easier/less effortful than a new activity because students already were given the correct response in a previous attempt.

For this purpose, we separated the counts of activities a student completed into two counts: number of activities the student completed only once (unique activities) and the number of activities the student completed more than once (repeated activities). We then repeated the same regression analyses as in the previous analyses (see Supplementary Materials for details) with activity completion separated by unique and repeated attempts. Note that although repeating activities was a low incidence event, it varied across modules (see Table 1), allowing for these analyses. The results of this analysis show that completing more unique activities was positively related to better quiz grades (m3b; see Figure 5) $\beta = 0.150, CI: [0.026, 0.271], \chi^2 = 5.69, p = .017, d = 0.06$, whereas completing more repeated activities had no statistically detectable relation to quiz grades (m3c) $\beta = -0.020, CI: [-0.044, 0.007], \chi^2 = 3.51, p = .061, d = -0.04$. There was not a statistically detectable interaction between unique and repeated activities on quiz grades (m3d) $\beta = -0.006, CI: [-0.034, 0.022], \chi^2 = 0.195, p = .659, d = -0.01$.

Figure 5

(a) Number of students-module pairs with different amounts of repeated and unique activities and (b) quiz performance as a function of repeated and unique activities for the psychology course (Study 1).



Note. The number of cases in each group in panel (a) along with the raw data for quiz grade are presented in panel (b) so the amount of data for each group can be considered. In addition, although the raw quiz grades are presented, other factors not represented in the plots (i.e., pretest scores) were considered in the regression analyses (see text for details).

In sum, the results of the analyses of student self-regulated study behavior in a psychology MOOC suggest that completing more activities is related to better learning outcomes. Moreover, students preferred the variability of completing different activities instead of repeating a previously completed activity. Completing more varied activities that cover a greater proportion of the module's learning objectives was particularly related with best learning outcomes. However, because exams were not tagged with learning objectives, we used whole exam grades as the outcome measure, making this analysis too coarse. The following study addresses this limitation and extends the findings to a different domain and population.

Study 2: Computing course

Overall, the results of the analyses of the Psychology course suggest that a practice testing effect is observed in natural contexts: completing more activities was associated with better learning outcomes. Follow up analyses suggested that this effect was not related to the repeated completion of the same activity but instead to the completion of more varied activities that cover more of the module's learning objectives.

In this second study, we evaluated the generalizability of these findings. We conducted the same set of analyses using a substantially different type of course. This course focused on a different domain (computation), had a different instructional format (blended learning), had different types of assessments (including pre and post quiz assessments before and after each module), a different population (included only students enrolled in a private higher education institution), was a requirement for students enrolled in a degree, and had a shorter duration (fewer weeks than the Psychology course).

This dataset also allows us to address two potential drawbacks of the Psychology dataset; namely, in the computing course we had access to question-by-question quiz results each tagged by learning objective, which allowed us to analyze the relation between practice, pretest, and quiz performance at the topic, or learning objective, level (instead of the module level as in Study 1).[2] Moreover, in this study we include per module pretest as a measure of module difficulty to account for variation in students' prior knowledge from module to module (instead of a single overall ability pretest as in Study 1).

Data and Methods

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study, and we follow JARS (Kazak, 2018). All data is available at Datashop.org (https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=2033 and https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=490) and all analysis code is available at GitHub ([removed to comply with masked review policy]). Data were analyzed using R, version 4.1.1 (R Core Team, 2021) and the following packages: lme4, version v1.1-14 (Bates et al., 2014), ggplot2, version 3.3.2 (Wickham, 2016), and EMAtools, version 0.1.3 (Kleiman, 2021). This study's design and its analysis were not pre-registered.

Course and Participants

In this study, we analyzed data from two sections of the course Computing at Carnegie Mellon (C@CM) from Fall 2010. C@CM is a three-unit, pass/fail mini-course. The goal of the course is to help students develop foundational computing and information literacy skills, focusing on the tools and technologies that are specific to Carnegie Mellon so students can be successful in other academic courses. The course comprises four primary modules: responsible computing, effective computing, safe computing, and information literacy. This course used a flipped-classroom approach where lectures, readings, and practice activities

[2] It was not possible to ascertain which parts of the text were associated with each learning objective and which parts of the text the student was reading. Thus, for all analyses comparing activity use with page access, comparisons were done at the module level and results for pre- and post- quizzes averaged across all questions.

were online, and weekly recitations were held with a teaching assistant for reviewing their work and getting clarifications.

A total of 1,512 students enrolled in the course and agreed to have their data analyzed and1,467 completed the course, that is, completed the final exam. Of these, we included in the analyses 703 students who completed the course (have final exam data) and for whom interaction data with the course was available. For detailed breakdown of the available data see table S2 in Supplementary information. Of note, students who did not complete the exam were unlikely to complete the quizzes and pretest quizzes, and some students who did complete both did not interact with the course, making it impossible to extract measures of reading and activity completion and were therefore excluded. No other exclusion criteria were used. As done in Study 1, we focused on students' use of the OLI textbook materials and how it relates to performance in the quizzes.

Assessments

Students were required to complete a pre-quiz before each module, a quiz after each module, and a final exam. Thus, there was a pretest and posttest for each course module.

The pre-quizzes and quizzes were self-paced, and each student could take them when they decided. Students could take pre-quizzes as many times as they wished but were only allowed to take each quiz once. The final exam was taken by all students during the same period. The content of the course was available during the exam time, but students must have completed all pre-quizzes and quizzes before being allowed to complete the exam.

Students completed all the assessments through OLI, and thus we had access to question-by-question results and each question was labeled with the target learning objective allowing for a more granular level of analyses.

Measures of student resource use

We extracted the same measures of student resource use that we did for the Psychology course. Although the content was different, the structure of the OLI textbook was like that of the Psychology course (see Table 2 for details of available resources in the Computing course).

Because the computing course included a pretest for each module, we could have used completion of the pretest as our measure student self-regulated testing. However, we opted to not do that for four main reasons. First, completing activities with feedback is potentially a very different instructional event than completing activities without feedback (e.g., Kang et al., 2007). Second, most students completed the pretest for most modules (less than 5% of the student-module pairs in the data did not include pretest). This lack of variability poses serious issues to our analyses. Third, if pretest completion is a measure of self-regulated testing, then it cannot also be a measure of prior knowledge, which is important to include for the reasons mentioned above. Fourth, maintaining the same measure as in Study 1 allows us to compare the two studies more directly.

Learning Objective tagging

Each activity, pretest question, and quiz question were mapped to learning objectives.

Data Analyses

We used the same approach as in Study 1 except for the following change. As noted above, in this course pretest/quiz questions were tagged with learning objectives. Consequently, the initial analyses of page vs. activity completion as well as coverage were done at the module level, whereas all other analyses were done at the learning objective level. The different analyses are because pages were not tagged with learning objectives and our inspection indicated that each page included more than one learning objective and it was not possible to ascertain which part of the page the student was spending their time on.

Results and discussion

Overall resource use

Table 3 includes the number of available resources, the percent of students who completed at least one of each resource available, and the average number of resources completed for each module. Most students completed at least one activity and one page across most modules. Although a smaller number of students completed repeated activities, there were students repeating activities across all modules.

Table 3

Descriptive Counts of Available Resources, Number of Students Who Completed At least One Resource, Average Number of Resources Completed, and Percentage of Resources Completed for Study 2.

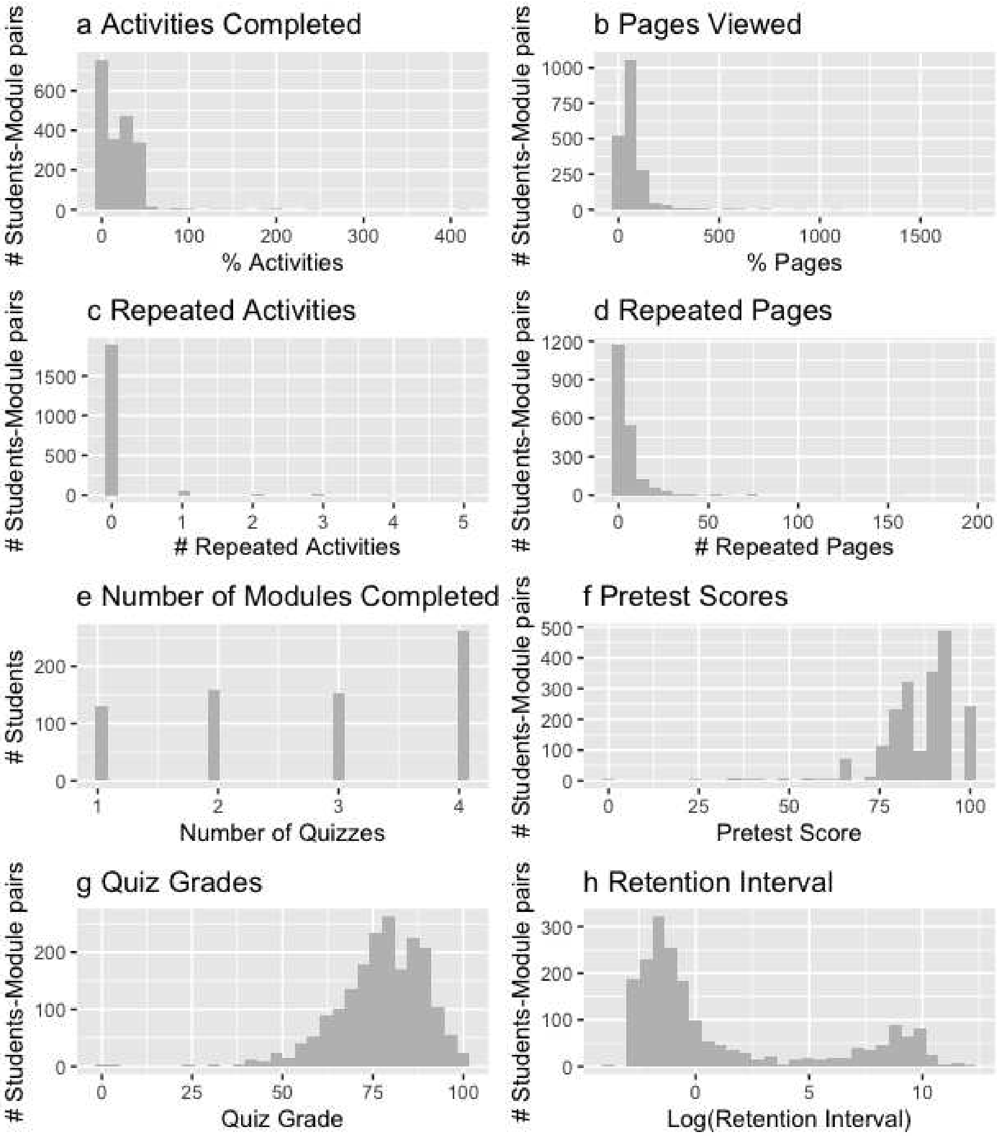
Module	Available Resources			% Students who completed at least one			Average number completed			Average % Completed		
	Activities	Pages	Learning Objectives	Activities	Repeated Activities	Pages	Activities	Repeated Activities	Pages	Activities	Pages	Learning Objectives
1	37	29	56	77%	4%	98%	7.75	0.06	17.57	21%	61%	20%
2	34	46	50	69%	6%	95%	6.47	0.08	24.34	19%	53%	26%
3	7	10	54	64%	0%	95%	1.41	0.00	9.68	20%	97%	12%
4	31	29	59	66%	2%	95%	4.26	0.04	16.69	14%	58%	25%

Note. N = 703. Available resources: counts of each type of resource available in the course. % students who completed at least one: percent of students (out of the total number of students available) who completed at least one of the resources available for that module. Average number completed: average number of resources completed for each module across all students; Average % completed: average (across students) percentage of available resources that were completed by student.

Initial inspection of the data indicated that students completed on average 18% of the available activities (*SEM* = 0.49, range = [0%,414%]) and 65% of the pages available (*SEM* = 1.79%, range = [0%,175%]. Students completed on average 5.34 (*SEM* = 0.13, range = [0, 41]) activities per module only once, with an average of 0.05 (*SEM* = 0.01, range = [0, 5]) repeated activities per module. Conversely, students studied an average of 12.62 (SEM = 0.26, range = [0, 120]) pages only once, with an average of 5.04 repeated pages (*SEM* = 0.22, range = [0, 196]). Mean performance was, on average, 87% on the pretest quizzes (*SEM* = 0.001%, range = [0%, 100%]) and 77% on quizzes (*SEM* = 0.01%, range = [0%, 100%]). Note that although the average grade on pretest quizzes is higher than on the quizzes, this reflects the fact that students were allowed to take the pretest as many times as they wished to achieve the score they wished. The posttest quiz could only be taken once. On average, students waited 46 minutes between finishing working on the textbook and taking the quiz (*SEM* = 3 minutes, range = [0, 1.63 days]). Visual inspection of these measures, as indicated in Figure 6, shows good distribution and variability.

Figure 6

Histograms showing distribution and variation of the different measures used for analysis for the computing course (Study 2)



Note. Counts of student-module pairs (i.e., number of data points) are presented for each predictor measure used in the regressions (activities completed, pages viewed, repeated activities, repeated pages, pretest score), and outcome measure used in the regression models (quiz grades). Number of quizzes completed by students is presented in panel e and retention interval (log transformed) is presented in panel h.

Like what we saw in the Psychology dataset, variability was higher across modules in the same student than across students in the same module for both activities and pages. Comparing how many activities students completed in a module with how many activities students completed in all other modules also shows results like the Psychology dataset (see Table 4).

Table 4

Frequency of student-module pairs by activity quartile for activities completed within a module or for all other modules (outside module) in the Computing Dataset (Study 2).

	Outside of Module Usage Quartile				
Module Usage Quartile	25-50	50-75	Bottom 25%	Top 25%	Module Totals
25-50	168	162	102	58	490
50-75	61	110	191	126	488
Bottom 25%	231	163	60	36	490
Top 25%	30	55	135	267	487
Outside Module Totals	490	490	488	487	1,955

Note. N = 703. Number of students-modules pairs (data points) in each quartile of activity usage for module activities (rows) and activity usage on other modules (columns). Data points in the diagonal line suggest consistent activity usage, that is, similar activity usage in the target module and other modules. Most students vary in how many activities they complete from module to module, as suggested by the variation in frequency away from the diagonal line.

Although the Computing course included fewer pages and activities and shorter temporal duration, we still see the same overall pattern as in the Psychology course: higher use of pages than activities and higher variability across modules than across students.

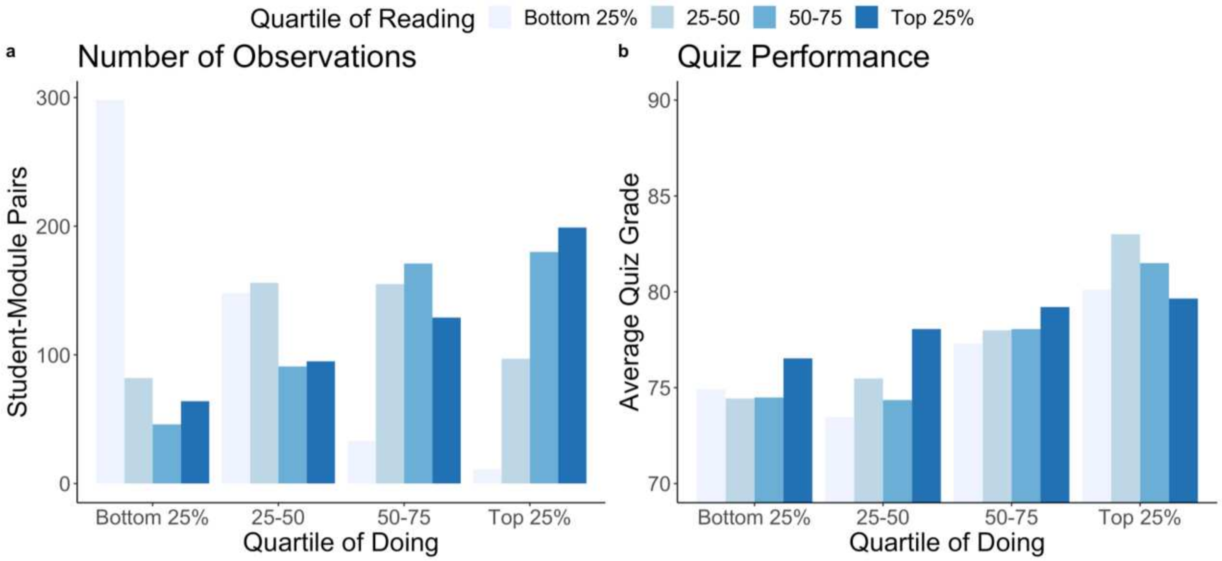
Dosage of practice and learning outcomes

Similarly, to what we saw for the psychology course, higher correctness in solving activities was related to higher quiz scores (m4b), $\beta = 0.06, CI: [-0.003, 0.093], \chi^2 = 35.16, p < .0001, d = 0.08$. Additionally, completing a larger percentage of activities was related to higher quiz scores, $\beta = 0.22, CI: [0.155, 0.279], \chi^2 = 24.39, p < .0001, d = 0.31$ (m4c), whereas completing more activities on the other modules was not (m4d), $\beta = 0.03, CI: [-0.026, 0.081], \chi^2 = 1.34, p = .247, d = 0.05$. Moreover, studying a larger percentage of pages was not related to changes in quiz scores, $\beta = -0.02, CI: [-0.061, 0.026], \chi^2 = 0.32, p = .571, d = -0.04$ (m4e) and neither was studying a larger percentage of pages on the other modules, $\beta = -0.04, CI: [-0.089, 0.002], \chi^2 = 1.88, p = .171, d = -0.07$. The interaction between completing more pages and completing more activities was also statistically detectable (m4g), $\beta = -0.04, CI: [-0.059, -0.024], \chi^2 = 22.89, p < .0001, d = -0.21$ (m4e; see Figure 7), suggesting that students benefited from completing more activities particularly when reading fewer pages and benefited from reading more pages particularly when completing fewer activities. That is, to perform well in the quizzes students must learn the content and if it is not done by completing activities (more efficient) it must be done by reading the text. A student cannot get a high quiz score without engaging with the material in some way.

These results replicate the findings of Study 1 suggesting that choosing to complete more activities is related to better quiz performance whereas choosing to complete more pages is not. And, as in Study 1, these results do not seem to be due to overall preferences to complete more pages. However, it also suggests that students learn from reading pages; when not completing activities, better quiz grades are associated with accessing more pages (and vice versa).

Figure 7

(a) Number of students-module pairs with different amounts of doing and reading and (b) quiz performance as a function of amount of doing and reading for the computing course (Study 2)



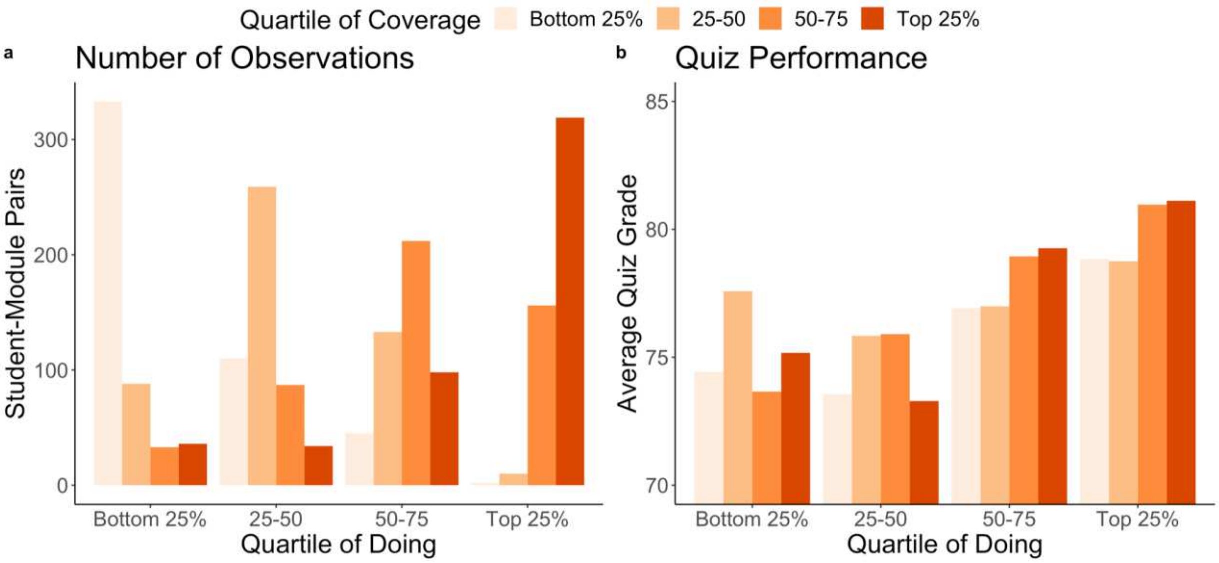
Note. The number of cases in each group in panel (a) along with the raw data for quiz grade are presented in panel (b) so the amount of data for each group can be considered. In addition, although the raw quiz grades are presented, other factors not represented in the plots (i.e., pretest scores) were considered in the regression analyses (see text for details).

Relation between learning objective coverage and quiz scores

We repeated the same analyses as for study 1 to investigate if completing more activities that cover more learning objectives was particularly related to best learning outcomes, as in Study 1. Completing activities related to more learning objectives was indeed related to better learning outcomes, $\beta = 0.02, CI: [-0.03, 0.08], \chi^2 = 3.91, p = .047, d = 0.04$ (m5b), though the effect size is small. Like what we saw for the psychology course, there was a statistically detectable interaction between coverage (number of learning objectives covered) and number of activities completed, $\beta = -0.05, CI: [-0.06, -0.03], \chi^2 = 21.25, p < .001, d = -0.21$ (m5c). Completing more activities was particularly related to better learning outcomes if the activities covered more learning objectives (see Figure 8).

Figure 8

(a) Number of students-module pairs with different amounts of doing and coverage and (b) quiz performance as a function of amount of doing and coverage for the computing course (Study 2).



Note. The number of cases in each group in panel (a) along with the raw data for quiz grade are presented in panel (b) so the amount of data for each group can be considered. In addition, although the raw quiz grades are presented, other factors not represented in the plots (i.e., pretest scores) were considered in the regression analyses (see text for details).

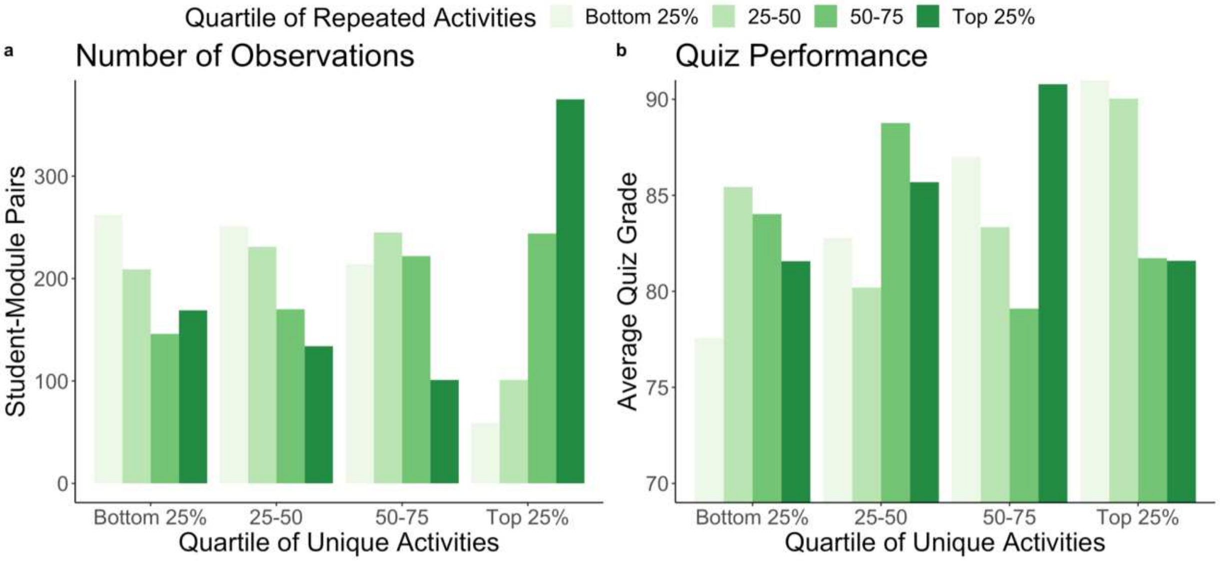
Relation between completing more different or the same activity more times and quiz outcomes

To further analyze the effect of coverage reported in the previous section, we investigated whether completing more unique or repeated activities for each learning objective influenced quiz performance on questions related to that same learning objective. To do that, we used the tagging of quiz, pretest, and activity questions for each learning objective and counted for each learning objective how many unique or repeated problems the student completed and their performance in the outcome measure.[3] Because these analyses include multiple measures for the same student at the learning objective level, we included student and learning objective as crossed random effects, unlike in previous analyses (which included student and module as random effects). Note that although repeating activities was a low incidence event, it varied across modules (see Table 3), allowing for these analyses. Completing more unique activities for a learning objective was related to better performance on quiz questions regarding that learning objective, $\beta = 0.06, CI: [0.03, 0.10], \chi^2 = 11.25, p < .001, d = 0.13$ (m7b), whereas completing more repeated activities was not, $\beta = -0.01, CI: [-0.03, 0.02], \chi^2 = 0.01, p = .913, d = -0.02$ (m7c). There was no statistically detectable interaction between the two variables, $\beta = -0.03, CI: [-0.07, 0.01], \chi^2 = 2.34, p = .126, d = -0.06$ (m7d; see Figure 9). Overall, the results of Study 2 replicate and extend the findings of Study 1.

[3] For completeness, we also analyzed data at the module level as in Study 1. We found the same pattern of results (see Supplementary Material).

Figure 9

(a) Number of students-module pairs with different amounts of repeated and unique activities and (b) quiz performance as a function of repeated and unique activities for the computing course (Study 2).



Note. The number of cases in each group in panel (a) along with the raw data for quiz grade are presented in panel (b) so the amount of data for each group can be considered. In addition, although the raw quiz grades are presented, other factors not represented in the plots (i.e., pretest scores) were considered in the regression analyses (see text for details).

General Discussion

The present work had two main goals: (1) an ecological valid assessment of the effect of self-regulated practice testing on learning outcomes and (2) evaluate whether exact repetition of practice problems is necessary for the effect.

By analyzing data from two courses, over 1,000 students, and multiple modules per student (for a total of more than 10,000 data points), across different semesters, student populations, academic years, and domains (computing and psychology), we showed—to the best of our knowledge for the first time—that self-regulated practice testing in a natural learning environment is associated with improved learning outcomes compared to viewing more pages.

Importantly, the association between the number of activities completed with quiz outcomes had medium effect sizes (Cohen *d*s between 0.25 and 0.31). Put another way, an increase of one activity completed was associated with improvements of 0.16 and 0.22 standard deviations on quiz scores, across the two studies. This result is equivalent to one percentage point increase in quiz grade per activity completed. Comparatively, viewing more pages was not associated with improvements in quiz scores, albeit as noted below, our measure of reading is potentially noisier than that of completing activities, which might contribute to this weak relation. Regardless, the relation between completing more activities and quiz grades was *five times larger* than the relation between viewing more pages and learning outcomes in Study1, and *11 times larger* in Study 2; even accounting for some noise in the measure, this are staggering differences.

Our results offer new insights into the theoretical underpinnings of why practice testing is related to better learning outcomes. Across both courses, we found a clear dosage effect: increased practice testing was related to increased assessment grades, when comparing outcomes for the same student across multiple modules of the same course. These results are consistent with retrieval-effort theories of practice testing (Glover, 1989), positing that more opportunities to retrieve information with an appropriate level of effort should result in better learning than equivalent opportunities to read the information. Although previous laboratory research has found similar results (e.g., Karpicke & Roediger, 2008; McDermott, 2006; Roediger & Karpicke, 2006b; Vaughn & Rawson, 2012), results from classroom studies have been mixed (e.g., Foss & Pirozzolo, 2017), and a comprehensive meta-analysis (Rowland, 2014) and a meta-analysis focused specifically on transfer (Pan & Rickard, 2018) did not show evidence for this type of effect. Thus, our analysis is the first to show large-scale naturalistic evidence of this dosage effect, providing some support for retrieval-effort theories of learning.

One possibility for why our results differ from those of the meta-analyses mentioned lies with the characteristics of the practice testing associated with best learning outcomes in our data. First, students overwhelmingly preferred to complete varied activities instead of repeating the same activity, which is not the traditional laboratory setup where repeating the same activity is common for practice testing. Moreover, we found that increasing varied activities—as opposed to exact repetition of the same activity—was particularly associated with better learning outcomes. These results are consistent with retrieval-effort theories of practice testing (Glover, 1989) under the assumption that repeating the same activity multiple times is inherently easier because the correct answer is part of the feedback than completing multiple different activities on the same content.

It is possible that the effect of repeating the same activity depends on the nature of the repetition and the outcome measure such that repeated tests with the same problem improve memory whereas repeated tests with different problems improves generalization. Thus, repeated *varied* tests (as most students completed in the current studies) could be particularly beneficial for transfer situations such as what is common in classroom contexts (and in the datasets analyzed here) where the student is asked to answer *different* questions about the *same* topic. Consistent with this explanation, Pan & Rickard's (2018) meta-analysis found that elaboration during testing practice—which included variation in the questions as well detailed feedback as present in the current study—was a statistically significant moderator of the effect. Unfortunately, the authors did not investigate relations between moderators, thus not being able to identify the interaction we propose. However, additional congruent evidence for such interaction comes from studies testing participant's memory for trained materials. For example, in a study using exact repetition of the same practice during training and memory tests as the learning outcome, Pyc and Rawson (2009) did find a dosage effect but with diminished returns. Thus, it is possible that even for memory tasks increasing repeated testing practice of the same materials could improve learning but to a point, which would help explain why some previous studies found such evidence, but the meta-analysis did not.

In sum, we see a dosage effect of testing that seems connected to increased varied tests and differs from findings from previous laboratory studies and meta-analyses. Our hypothesis, congruent with previous evidence, is that this finding is related to the nature of the outcome tests (transfer rather than memory) and the repetition during practice tests (varied tests as opposed to the same test). This novel hypothesis generated from analyses of educational big data can be further tested using laboratory and classroom studies.

More broadly, not finding a positive relation between repeatedly practicing the same activity and outcome measures raises the question of whether the benefit of practice testing is connected to retrieving previously encoded information during study as some theories propose (Benjamin & Tullis, 2010; Carpenter, 2009), or changing how information is encoded the first time it is presented (Pavlik & Anderson, 2008; Pyc & Rawson, 2009). If the benefits of practice testing were primarily associated with repeated retrieval of the same information, a benefit of exact repetition should be seen–as it increases the number of exact retrievals. Yet, we saw a decreasing association between repeated tests and performance. Thus, our findings weigh against a need for repetition of identical testing questions and support for the hypothesis that the effect of testing is due to changes in how information is encoded. Compared to reading a text where all the information is provided, completing an activity requires students to generate a response and compare it with the feedback, if provided. One possibility is that more effortful retrievals that do not come from repeating the same question lead to better encoding and slower forgetting (Pavlik & Anderson, 2008; Pyc & Rawson, 2009) in addition to increasing the routes to successful retrieval compared to reading or repeating the same activity multiple times (McDaniel and Masson, 1985). Importantly, although retention interval has been shown to influence the impact of retrieval practice (Roediger & Karpicke

2006b; but see Runquist, 1983), we did not analyze such an interaction. On average students waited around an hour after finishing working on the materials to start the quiz, but the true retention interval between finishing working on an activity and completing a related question in the quiz is a complex measure to derive from the present data. Regardless, it is possible that the findings presented are in part due to the retention interval students chose.

Our results seem to indicate that inducing skills/concepts that function across a variety of task activities is harmed by lack of variability in tests (Koedinger et al., 2012) and are consistent with theories of concept acquisition suggesting that best transfer and generalization is achieved by studying varied examples to promote extraction of common features or schemas (e.g., Catrambone & Holyoak, 1989; Gick & Holyoak, 1980; Paas et al., 1994). Our results are consistent with the hypothesis that the benefit of practice testing is the result of changes at encoding (by elaborating the same concept/skills across multiple, slightly different activities), rather than differences in how information is retrieved during study itself. That is, different tests on the same concept promote learning, compared to retrieving the same information with the same test multiple times. For example, learning to find the area of an irregular shape benefits more from completing multiple different problems with different shapes and values, than repeating the same problem multiple times. Similarly, learning that the plural of words ending in “-y” in English is often obtained by replacing “-y” with “-ies,” is better learned by practicing with multiple words than repeating practice of the same word.

More broadly, the current results speak to some limitations of students’ self-regulated learning. Although previous research has shown that self-paced (Tullis & Benjamin, 2011), self-ordered (Carvalho et al., 2016), and self-spaced (Ciccone & Brelsford, 1976) practice can yield better learning outcomes than when the practice is chosen for the students, it is clear that not all students make the right decision, and a self-regulated decision not to practice might yield worse results.

An important question not addressed by the current studies is why and when students decide to complete more activities or read more pages and repeat the same activity multiple times. One possibility is that students complete more activities/repeat the same activity more often when the content is more difficult. Although we found no effect of performance in the activities on the number of activities or pages completed, it is still possible that our measure of perceived difficulty is not capturing the critical aspects. It is also possible that students complete more activities for modules they are more interested in (although the same could be said for reading more pages). Even though the approach used here lacks the counterfactual to rule out any of these possibilities, previous research showing that students benefit more from completing more activities for the same content, even when they state preferring fewer activities, suggests that it is the fact that students completed more activities and not why they completed them that matters (Schnackenberg et al., 1998; Schnackenberg & Sullivan, 2000).

Big data, internal validity, and external validity

Although we took care in our analytical approach to reduce the potential alternative explanations of the findings, including using a within-subject analytic approach and explicitly ruling out potential overall ability/interest variables, it is possible that some other variable accounts for the results. Because we used a measure of pretest and compared student behavior across modules, it is unlikely that the effect is the result of different general characteristics of the students. However, it is still possible that differential variations in level of interest or motivation in particular topics can be the basis for the results observed. Thus, care is necessary to interpret the results as correlational evidence, but not causal prediction.

Learning by completing practice tests as described here involved effortful (Roediger & Karpicke, 2006a), active engagement and knowledge manipulation by the student (Wieman, 2014), with timely (Roediger & Karpicke, 2006a) and explanatory (Clark & Mayer, 2008) feedback. All these properties have been associated with better learning outcomes compared to passive learning situations such as reading. Any of these factors might have contributed to the benefits of completing more practice activities. Thus, although the research presented here does not provide ideal *internal validity* for causal links, it does enhance the *external validity* of evidence for the claims towards the benefits of study practice. Currently, evidence for a benefit of practice testing relies much more on internal validity than external validity, so our studies are an important addition.

This type of research might be particularly powerful for studying self-regulated learning. It might be impossible to randomly-assign students to do what is required from a condition, therefore self-regulated learning might always have to be studied observationally like what we have done here. However, self-regulated learning is one of the critical aspects of current education and its study–even if observational in nature–can inform theory and practice. If students' learning decisions are appropriately calibrated, and the decision itself changes the learning process (Gureckis & Markant, 2012), then forcing students to adopt a particular approach might have detrimental consequences instead of positive outcomes because the approach selected might not match how students approach the task (Carvalho et al., 2018). For example, Carvalho and collaborators (2016) showed in a yoked design that when students were allowed to self-regulate their learning by selecting how to sequence their study, the sequence of study that resulted in improved learning was not equally successful when yoked students were required to follow it. Similarly, using secondary analyses of existing data, Carvalho, Sana, and Yan (2020) showed that students who benefited the most from spacing their study were not the ones doing it as much.

Limitations

Secondary data analyses of existing detailed educational data offer excellent (and low-cost) opportunities to study well-established phenomena "in the wild." The purpose of this research was not to establish a causal relationship between practice and learning–many laboratory studies have done so before. Rather, its purpose was to explore whether such relations are observed in naturalistic settings where many variables in addition to the independent variable of interest vary randomly. In addition to the absence of direct causal links, the type of data available, the processing decisions, and how operationalized the variables of interest might have had an impact on the results. For instance, although care was taken not to double-count activity counts and page views, because activities were embedded in pages it is possible that some of the page views are in fact activity completions (but not the opposite as each interaction with an activity yielded a clear timestamp). However, this drawback is likely similar to what is the case in experimental situations where participants are asked to read but might just glance at a page instead. Similarly, at the other extreme, it is possible that students glanced at the text while completing an activity on the same page, and such reading was not counted.

Importantly, our analytic approach and careful consideration of alternative hypotheses provides strength even in the case of this potentially unaccounted for variability. First, we included “control” measures such as percentage of activities completed on other modules and percentage of reading on other modules and interaction terms between completing activities and reading. If our measures were too noisy or invalid, then there would be no reason to expect that percentage of activities completed or pages read in the module would vary inside and outside each module or be predictive of quiz outcomes; yet we found critical differences. Second, with limitations on measuring reading in mind, we focused our follow up analyses on the reasons why deciding to complete more activities might be related to better learning outcomes and not on the comparison with reading (though, as we noted above, the limitations with our measure of reading are likely similar to limitations of other measures of reading even in experimental work). Third, we considered alternative measures of the same construct, such as using pretest completion as a measure of self-regulated testing. Although the variability was very low — most students completed the pretest and did so a small number of times — not allowing for a full analysis, completing the pretest, and completing more activities were correlated, $r^2 = .10$ $p < .0001$.

Finally, one of the powers of analyses of existing data, in addition to its external validity mentioned above, is the ability to identify and generate novel hypothesis. The critical insight from the current work is that best learning is associated with varied practice with feedback, not with reading text. Indeed, recent classroom (Carvalho, Manke, & Koedinger, 2018) and laboratory studies (Carvalho, Sana, & Koedinger, 2021), both with random assignment and tight controls, further confirm that best learning outcomes are achieved when learners complete more activities, even when they do not read any relevant text.

Conclusion

There has been great interest and progress in techniques that improve learning and the underlying cognitive processes (Bjork et al., 2013; Dunlosky et al., 2013). However, much of this progress is based on small scale, tightly controlled, laboratory studies with limited scope of learning content. Although these studies allow us to carefully test hypotheses and establish causal relationships, one can well wonder how much external validity should be sacrificed for tight control and internal validity. Results of recent randomized controlled trials mentioned in the introduction heighten such concerns (Gurung & Burns, 2019). This question is particularly important when we try to predict learning where it matters: in natural contexts such as the classroom. In these contexts, many variables such as students' decisions, variations in control conditions (e.g., use of worked examples and self-explanation instead of passive reading), student background knowledge, differences in learning rate, etc., are likely to interact. We argue that to fully understand the cognitive processes involved in learning we need to advance the literature by investigating our hypotheses in natural educational contexts, both using controlled experimentation that emphasizes internal validity (Koedinger et al., 2010, 2012) and data science approaches with naturalistic data that emphasize external validity (Koedinger et al., 2015). The work presented here demonstrates one way to leverage data from naturally occurring datasets to inform theoretical developments and the understanding of cognitive phenomena. It contributes to practice and theoretical development by suggesting that self-regulated practice can improve long-term learning, particularly when it is effortful and accurate, providing evidence for some learning theories in natural environments.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*, *87*(3), 659–701.

Agarwal, P. K. (2019). Retrieval practice & Bloom’s taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, *111*(2), 189–209.

Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, *26*(2), 147–179.

Andergassen, M., Mödritscher, F., & Neumann, G. (2014). Practice and Repetition during Exam Preparation in Blended Learning Courses: Correlations with Learning Results. In *Journal of Learning Analytics* (Vol. 1, Issue 1, pp. 48–74). https://doi.org/10.18608/jla.2014.11.4

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology. Applied*, *24*(1), 43–56.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. In *arXiv [stat.CO]*. arXiv. http://arxiv.org/abs/1406.5823

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*(3), 228–247.

Bier, N., Lovett, M., & Seacord, R. (2011). An online learning approach to information systems security education. *Proceedings of the 15th Colloquium for Information Systems Security Education*, 56–62.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(5), 1118–1133.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(6), 1563–1569.

Carvalho, P. F., Braithwaite, D. W., de Leeuw, J. R., Motz, B. A., & Goldstone, R. L. (2016). An In Vivo Study of Self-Regulated Study Sequencing in Introductory Psychology Courses. *PloS One*, *11*(3), e0152115.

Carvalho, P. F., Gao, M., Motz, B. A., & Koedinger, K. R. (2018). *Analyzing the relative learning benefits of completing required activities and optional readings in online courses*.

Carvalho, P.F., Manke, K.J, & Koedinger, K.R. (2018). Not all Active Learning is Equal: Predicting and Explaining Improves Transfer Relative to Answering Practice Questions. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1458-1463). Austin, TX: Cognitive Science Society.

Carvalho, P.F., Sana, F. & Koedinger, K. R. (2021). Can Pre-Testing with Feedback Make Reading Text Unnecessary? Manuscript under review.

Carvalho, P. F., Sana, F., & Yan, V. X. (2020). Self-regulated spacing in a massive open online course is related to better learning. *NPJ Science of Learning*, *5*, 2.

Castro, R. M., Kalish, C., Nowak, R., Qian, R., Rogers, T., & Zhu, J. (2009). Human Active Learning. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21* (pp. 241–248). Curran Associates, Inc.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. In *Journal of Experimental Psychology: Learning, Memory, and Cognition* (Vol. 15, Issue 6, pp. 1147–1156). https://doi.org/10.1037/0278-7393.15.6.1147

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*, 73–105. doi:10.1111/j.1756-8765.2008 .01005.x

Ciccone, D. S., & Brelsford, J. W. (1976). Spacing repetitions in paired-associate learning: Experimenter versus subject control. In *Journal of Experimental Psychology: Human Learning and Memory* (Vol. 2, Issue 4, pp. 446–455). https://doi.org/10.1037/0278-7393.2.4.446

Clark, R. C., & Mayer, R. E. (2008). Learning by viewing versus learning by doing: Evidence-based guidelines for principled learning environments. *Performance Improvement*, *47*(9), 5–13.

Deans for Impact. (2015). *The Science of Learning*. Deans for Impact.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *14*(1), 4–58.

Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology*, *109*(8), 1067–1083.

Gick, M. L., & Holyoak, J. (1980). Analogical Problem Solving. *Cognitive Psychology*, *12*, 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. In *Cognitive Psychology* (Vol. 15, Issue 1, pp. 1–38). https://doi.org/10.1016/0010-0285(83)90002-6

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. In *Journal of Educational Psychology* (Vol. 81, Issue 3, pp. 392–399). https://doi.org/10.1037/0022-0663.81.3.392

Gureckis, T. M., & Markant, D. B. (2012). Self-Directed Learning: A Cognitive and Computational Perspective. *Perspectives on Psychological Science: A Journal of the*

Association for Psychological Science, *7*(5), 464–481.

Gurung, R. A. R., & Burns, K. (2019). Putting evidence- based claims to the test: A multi- site classroom study of retrieval practice and spaced practice. *Applied Cognitive Psychology*, *33*(5), 732–743.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*(4), 332–340.

Janes, J. L., Dunlosky, J., & Rawson, K. A. (2018). How Do Students Use Self-Testing Across Multiple Study Sessions When Preparing for a High-Stakes Exam? *Journal of Applied Research in Memory and Cognition*, *7*(2), 230–240.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the Test…or Testing to Teach: Exams Requiring Higher Order Thinking Skills Encourage Greater Conceptual Understanding. *Educational Psychology Review*, *26*(2), 307–329.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *The European Journal of Cognitive Psychology*, *19*(4-5), 528–558.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science, 331*(6018), 772–775.

Karpicke, J. D., & Roediger, H. L., 3rd. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968.

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist, 73*(1), 1-2. http://dx.doi.org/10.1037/amp0000263

Kim, A. S. N., Wong-Kee-You, A. M. B., Wiseheart, M., & Rosenbaum, R. S. (2019). The spacing effect stands up to big data. *Behavior Research Methods*, *51*(4), 1485–1497.

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education*, *104*, 18–33.

Kleiman, E. M. (2017). EMAtools: Data management tools for real-time monitoring/ecological momentary assessment data. Retrieved from https://cran.r-project.org/package=EMAtools

Koedinger, K. R., & Aleven, V. (2007). Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. *Educational Psychology Review*, *19*(3), 239–264.

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In *Handbook of educational data mining*.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, *36*(5), 757–798.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. *Wiley Interdisciplinary Reviews. Cognitive Science*, *6*(4), 333–353.

Liyanagunawardena, T. R., Parslow, P. and Williams, S. (2014) Dropout: MOOC participants' perspective. In: EMOOCs 2014, the Second MOOC European Stakeholders Summit, 1012 th February 2014, Lausanne, Switzerland, pp. 95-100. Available at http://centaur.reading.ac.uk/36002/

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3 Pt 1), 1242–1255.

Lovett, M., Meyer, O., & Thille, C. (2008). JIME - The Open Learning Initiative: Measuring the Effectiveness of the OLI Statistics Course in Accelerating Student Learning. In *Journal of Interactive Media in Education* (Vol. 2008, Issue 1, p. 13). https://doi.org/10.5334/2008-14

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 371–385.

https://doi.org/10/fjf96z

McDermott, K. B. (2006). Paradoxical effects of testing: repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition*, *34*(2), 261–267.

Nguyen, K., & McDaniel, M. A. (2015). Using Quizzing to Assist Student Learning in the Classroom: The Good, the Bad, and the Ugly. *Teaching of Psychology* , *42*(1), 87–92.

Paas, F. G. W. C., Fred G W, & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. In *Journal of Educational Psychology* (Vol. 86, Issue 1, pp. 122–133). https://doi.org/10.1037/0022-0663.86.1.122

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: choices and consequences. *Psychonomic Bulletin & Review*, *14*(2), 187–193.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology. Applied*, *14*(2), 101–117.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria. https://www.R-project.org/

Ridgeway, K., Mozer, M. C., & Bowles, A. R. (2017). Forgetting of Foreign-Language Skills: A Corpus-Based Analysis of Online Tutoring Software. *Cognitive Science*, *41*(4), 924–949.

Roediger, H. L., 3rd, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

Roediger, H. L., 3rd, & Karpicke, J. D. (2006a). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *1*(3), 181–210.

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Rothkopth, E. Z. (1968). Textual constraints as function of repeated inspection. *Journal of Educational Psychology, 59*(1), 20–25.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463.

Schnackenberg, H. L., & Sullivan, H. J. (2000). Learner control over full and lean computer-based instruction under differing ability levels. *Educational Technology Research and Development: ETR & D*, *48*(2), 19–35.

Schnackenberg, H. L., Sullivan, H. J., Leader, L. F., & Jones, E. E. K. (1998). Learner preferences and achievement under differing amounts of learner practice. In *Educational Technology Research and Development* (Vol. 46, Issue 2, pp. 5–16). https://doi.org/10.1007/bf02299786

Son, L. K., & Kornell, N. (2009). Simultaneous decisions at study: time allocation, ordering, and spacing. *Metacognition and Learning*, *4*(3), 237–248.

Stafford, T., & Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological Science*, *25*(2), 511–518.

Steyvers, M., & Benjamin, A. S. (2019). The joint contribution of participation and performance to learning functions: Exploring the effects of age in large-scale data sets. *Behavior Research Methods*, *51*(4), 1531–1543.

Toppino, T. C., LaVan, M. H., & Iaconelli, R. T. (2018). Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying versus retrieval practice. *Memory &*

Cognition, *46*(7), 1164–1177.

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language*, *64*(2), 109–118.

Tullis, J. G., Fiechter, J. L., & Benjamin, A. S. (2018). The efficacy of learners' testing choices. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(4), 540–552.

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review*, *19*(5), 899–905.

Wickham, H. (2016). ggplot2: elegant graphics for data analysis. Springer.

Wieman, C. E. (2014). Large-scale comparison of science teaching methods sends clear message [Review of *Large-scale comparison of science teaching methods sends clear message*]. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8319–8320.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory* , *20*(6), 568–579.

Wissman, K. T., Zamary, A., & Rawson, K. A. (2018). When Does Practice Testing Promote Transfer on Deductive Reasoning Tasks? *Journal of Applied Research in Memory and Cognition*, *7*(3), 398–411.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*(3), 214–221.

Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology. Applied*, *23*(3), 263–277.

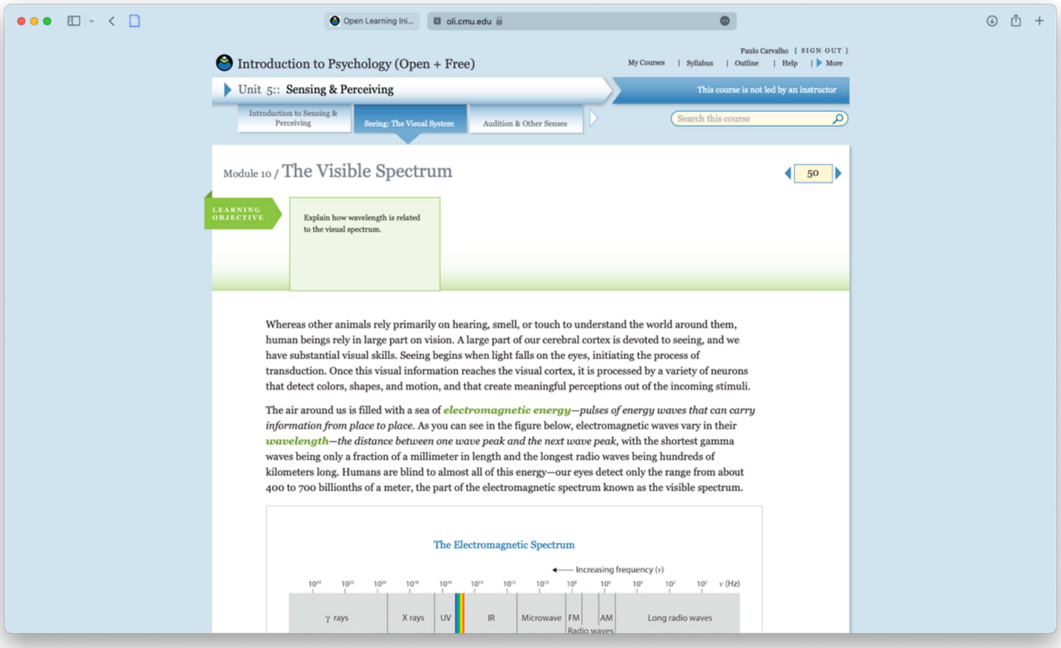
Supplementary Material

This supplementary material includes screenshots of the courses for the Psychology Course (Study 1) and the Computing Course (Study 2), detailed information about the samples (Table S1 for Study 1 and Table S2 for Study2), and detailed results of all the regression analyses in all studies (Tables S3 to S9 for Study 1 and Tables S10 to S17 for Study 2). Regression models are referred to by numbers matching the reference in the main manuscript. R code is presented as reference.

Examples of the materials

Figure S1

Screenshots of the Psychology Course



Introduction to Psychology (Open + Free)

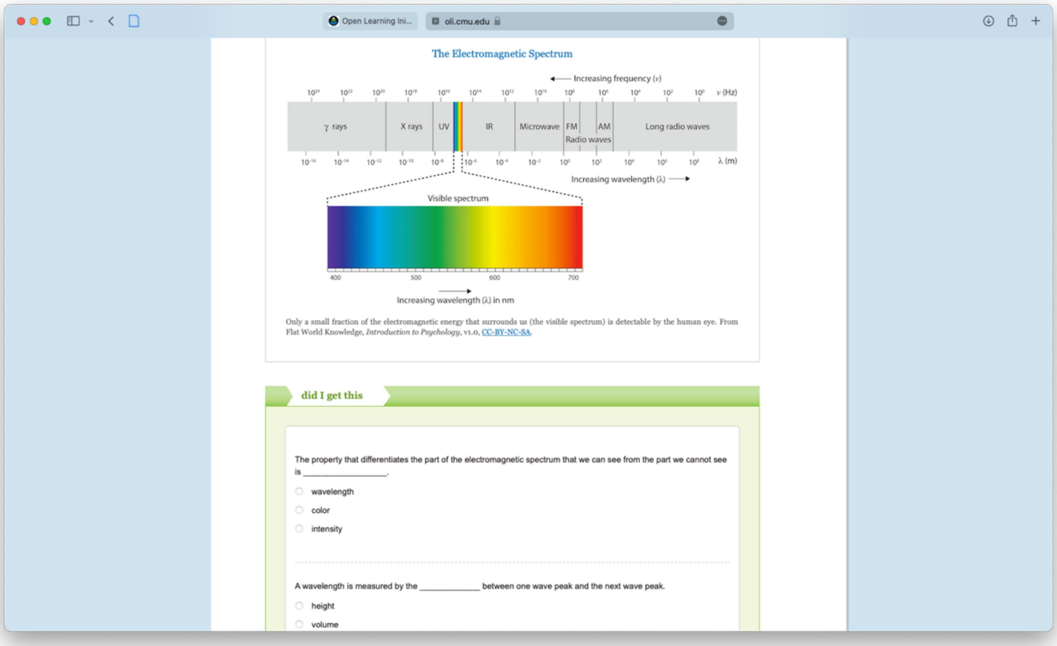
Module 10 / The Visible Spectrum

Explain how wavelength is related to the visual spectrum.

Whereas other animals rely primarily on hearing, smell, or touch to understand the world around them, human beings rely in large part on vision. A large part of our cerebral cortex is devoted to seeing, and we have substantial visual skills. Seeing begins when light falls on the eyes, initiating the process of transduction. Once this visual information reaches the visual cortex, it is processed by a variety of neurons that detect colors, shapes, and motion, and that create meaningful perceptions out of the incoming stimuli.

The air around us is filled with a sea of ***electromagnetic energy****—pulses of energy waves that can carry information from place to place*. As you can see in the figure below, electromagnetic waves vary in their ***wavelength****—the distance between one wave peak and the next wave peak*, with the shortest gamma waves being only a fraction of a millimeter in length and the longest radio waves being hundreds of kilometers long. Humans are blind to almost all of this energy—our eyes detect only the range from about 400 to 700 billionths of a meter, the part of the electromagnetic spectrum known as the visible spectrum.

The Electromagnetic Spectrum



The Electromagnetic Spectrum

Only a small fraction of the electromagnetic energy that surrounds us (the visible spectrum) is detectable by the human eye. From Flat World Knowledge, *Introduction to Psychology*, v1.0, CC-BY-NC-SA.

did I get this

The property that differentiates the part of the electromagnetic spectrum that we can see from the part we cannot see is ________________.

○ wavelength

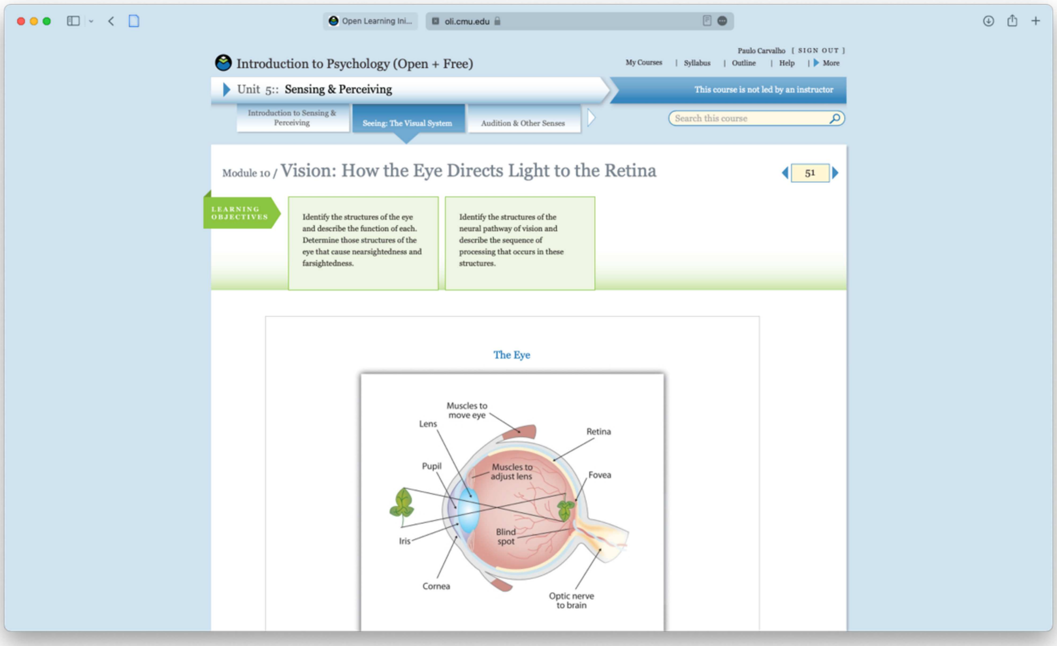
○ color

○ intensity

A wavelength is measured by the ____________ between one wave peak and the next wave peak.

○ height

○ volume



Introduction to Psychology (Open + Free)

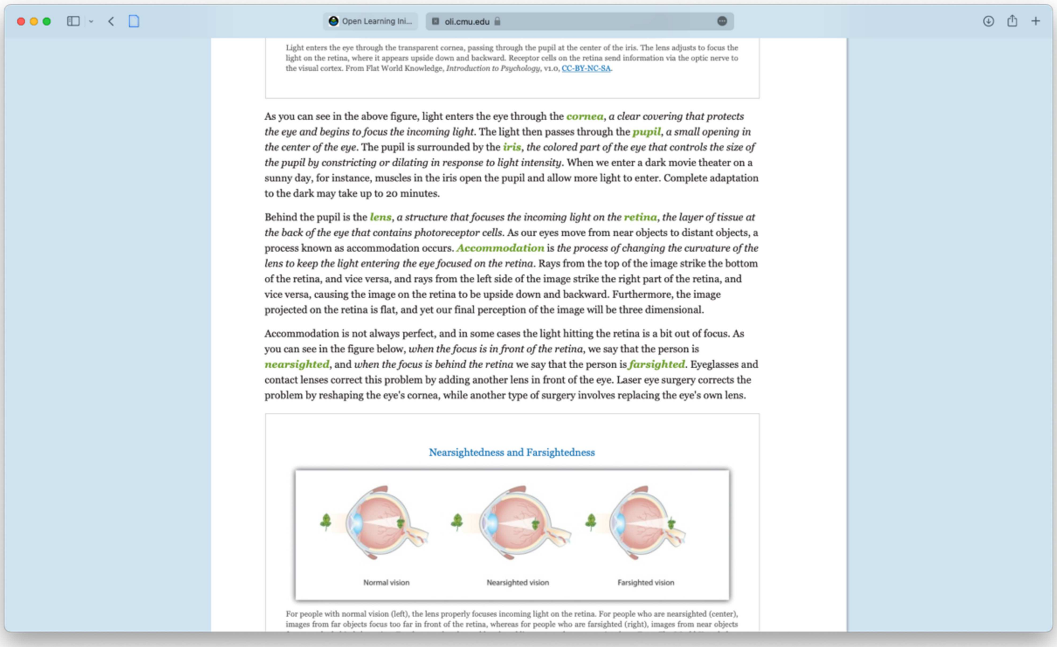
Unit 5:: Sensing & Perceiving

Module 10 / Vision: How the Eye Directs Light to the Retina

Identify the structures of the eye and describe the function of each. Determine those structures of the eye that cause nearsightedness and farsightedness.

Identify the structures of the neural pathway of vision and describe the sequence of processing that occurs in these structures.

The Eye



Light enters the eye through the transparent cornea, passing through the pupil at the center of the iris. The lens adjusts to focus the light on the retina, where it appears upside down and backward. Receptor cells on the retina send information via the optic nerve to the visual cortex. From Flat World Knowledge, *Introduction to Psychology*, v1.0, CC-BY-NC-SA.

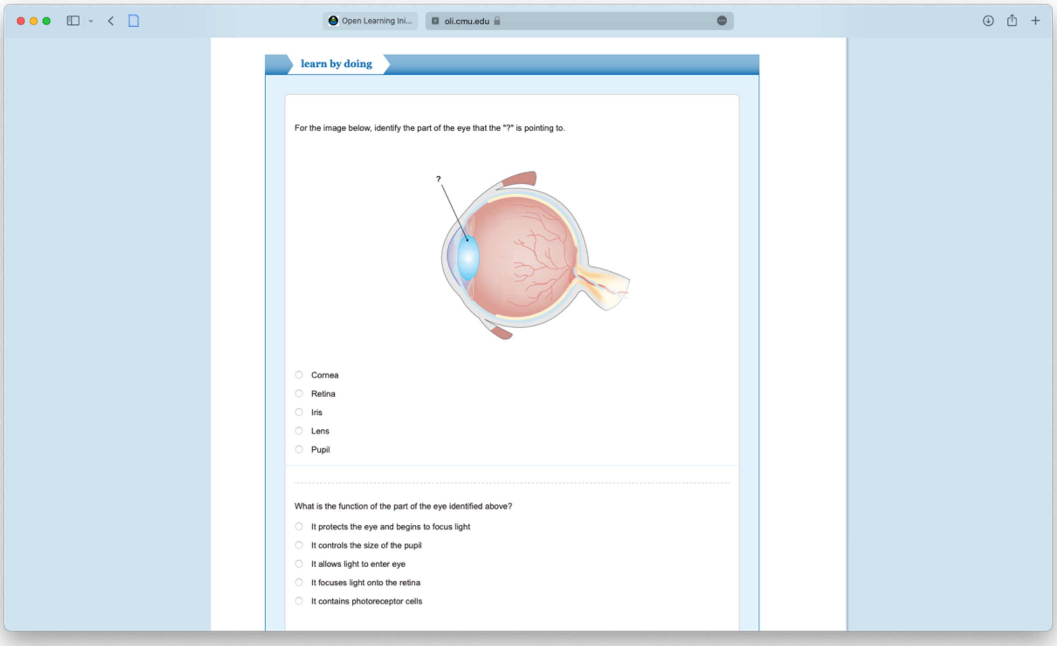
As you can see in the above figure, light enters the eye through the ***cornea***, *a clear covering that protects the eye and begins to focus the incoming light*. The light then passes through the ***pupil***, *a small opening in the center of the eye*. The pupil is surrounded by the ***iris***, *the colored part of the eye that controls the size of the pupil by constricting or dilating in response to light intensity*. When we enter a dark movie theater on a sunny day, for instance, muscles in the iris open the pupil and allow more light to enter. Complete adaptation to the dark may take up to 20 minutes.

Behind the pupil is the ***lens***, *a structure that focuses the incoming light on the* ***retina***, *the layer of tissue at the back of the eye that contains photoreceptor cells*. As our eyes move from near objects to distant objects, a process known as accommodation occurs. ***Accommodation*** is *the process of changing the curvature of the lens to keep the light entering the eye focused on the retina*. Rays from the top of the image strike the bottom of the retina, and vice versa, and rays from the left side of the image strike the right part of the retina, and vice versa, causing the image on the retina to be upside down and backward. Furthermore, the image projected on the retina is flat, and yet our final perception of the image will be three dimensional.

Accommodation is not always perfect, and in some cases the light hitting the retina is a bit out of focus. As you can see in the figure below, *when the focus is in front of the retina*, we say that the person is ***nearsighted***, and *when the focus is behind the retina* we say that the person is ***farsighted***. Eyeglasses and contact lenses correct this problem by adding another lens in front of the eye. Laser eye surgery corrects the problem by reshaping the eye's cornea, while another type of surgery involves replacing the eye's own lens.

Nearsightedness and Farsightedness

For people with normal vision (left), the lens properly focuses incoming light on the retina. For people who are nearsighted (center), images from far objects focus too far in front of the retina, whereas for people who are farsighted (right), images from near objects

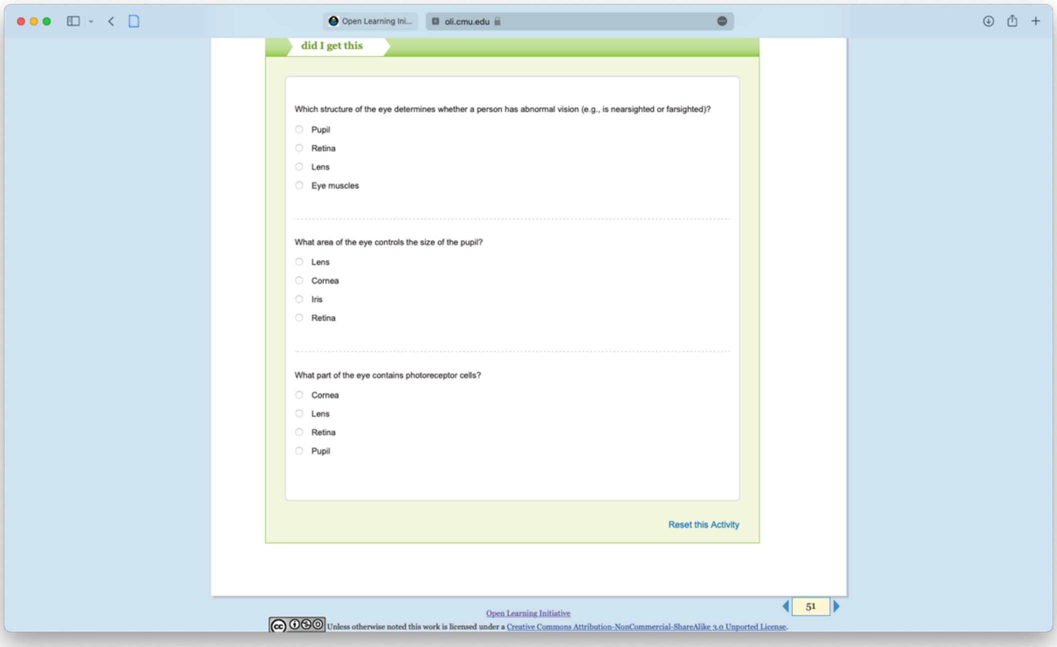


For the image below, identify the part of the eye that the "?" is pointing to.

○ Cornea
○ Retina
○ Iris
○ Lens
○ Pupil

What is the function of the part of the eye identified above?

○ It protects the eye and begins to focus light
○ It controls the size of the pupil
○ It allows light to enter eye
○ It focuses light onto the retina
○ It contains photoreceptor cells



Which structure of the eye determines whether a person has abnormal vision (e.g., is nearsighted or farsighted)?

○ Pupil
○ Retina
○ Lens
○ Eye muscles

What area of the eye controls the size of the pupil?

○ Lens
○ Cornea
○ Iris
○ Retina

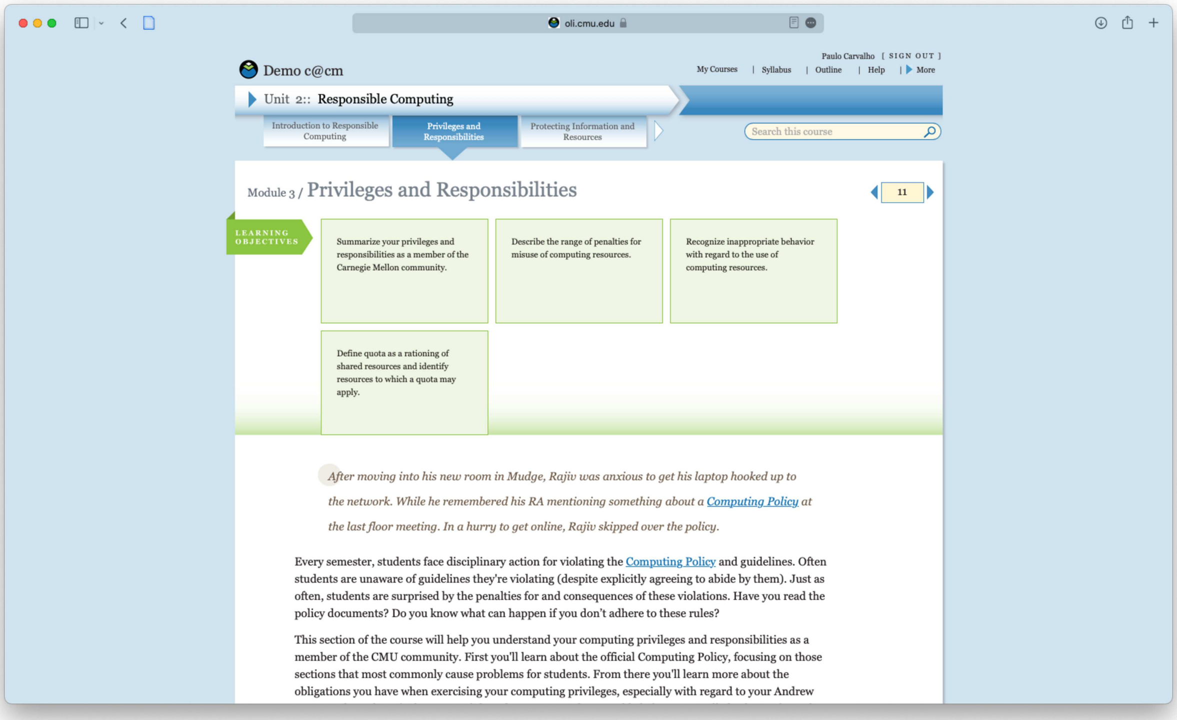
What part of the eye contains photoreceptor cells?

○ Cornea
○ Lens
○ Retina
○ Pupil

Open Learning Initiative

Figure S2

Screenshots of the Computing Course


oli.cmu.edu
Paulo Carvalho [SIGN OUT]
Demo c@cm
My Courses | Syllabus | Outline | Help | More
Unit 2:: Responsible Computing
Introduction to Responsible Computing
Privileges and Responsibilities
Protecting Information and Resources
Search this course
11
LEARNING OBJECTIVES

Module 3 / Privileges and Responsibilities

Summarize your privileges and responsibilities as a member of the Carnegie Mellon community.

Describe the range of penalties for misuse of computing resources.

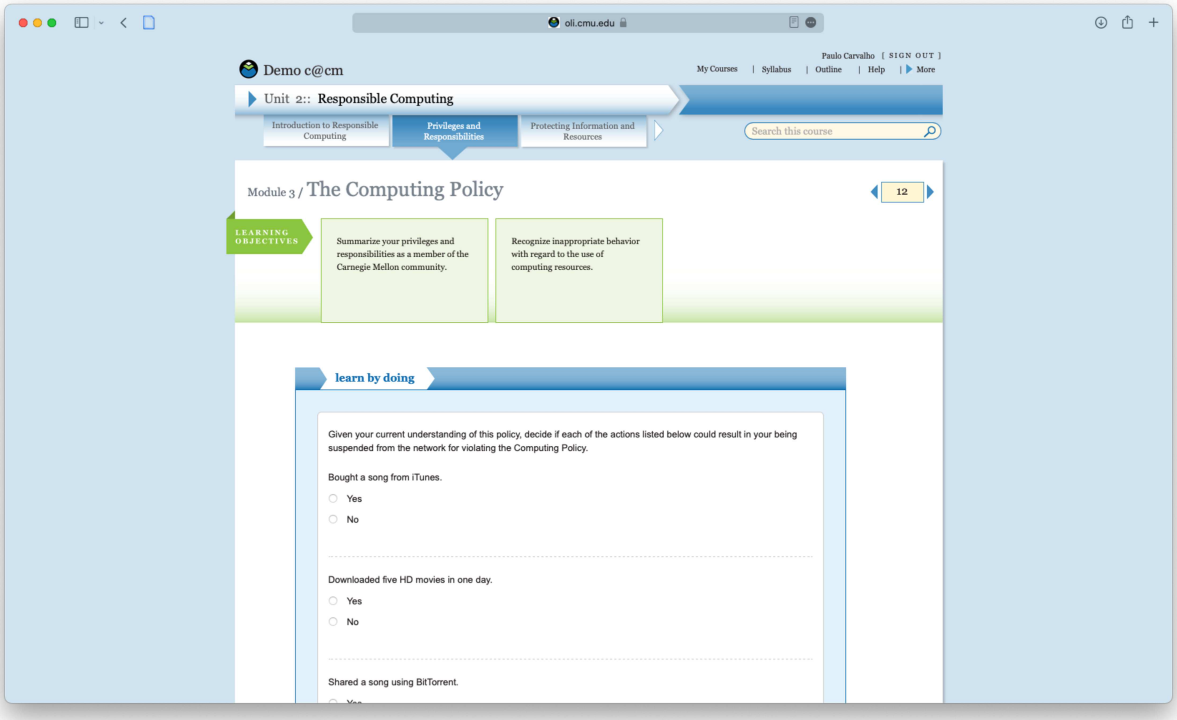
Recognize inappropriate behavior with regard to the use of computing resources.

Define quota as a rationing of shared resources and identify resources to which a quota may apply.

After moving into his new room in Mudge, Rajiv was anxious to get his laptop hooked up to the network. While he remembered his RA mentioning something about a Computing Policy at the last floor meeting. In a hurry to get online, Rajiv skipped over the policy.

Every semester, students face disciplinary action for violating the Computing Policy and guidelines. Often students are unaware of guidelines they're violating (despite explicitly agreeing to abide by them). Just as often, students are surprised by the penalties for and consequences of these violations. Have you read the policy documents? Do you know what can happen if you don't adhere to these rules?

This section of the course will help you understand your computing privileges and responsibilities as a member of the CMU community. First you'll learn about the official Computing Policy, focusing on those sections that most commonly cause problems for students. From there you'll learn more about the obligations you have when exercising your computing privileges, especially with regard to your Andrew


oli.cmu.edu
Paulo Carvalho [SIGN OUT]
Demo c@cm
My Courses | Syllabus | Outline | Help | More
Unit 2:: Responsible Computing
Introduction to Responsible Computing
Privileges and Responsibilities
Protecting Information and Resources
Search this course
12
LEARNING OBJECTIVES
learn by doing

Module 3 / The Computing Policy

Summarize your privileges and responsibilities as a member of the Carnegie Mellon community.

Recognize inappropriate behavior with regard to the use of computing resources.

Given your current understanding of this policy, decide if each of the actions listed below could result in your being suspended from the network for violating the Computing Policy.

Bought a song from iTunes.

○ Yes

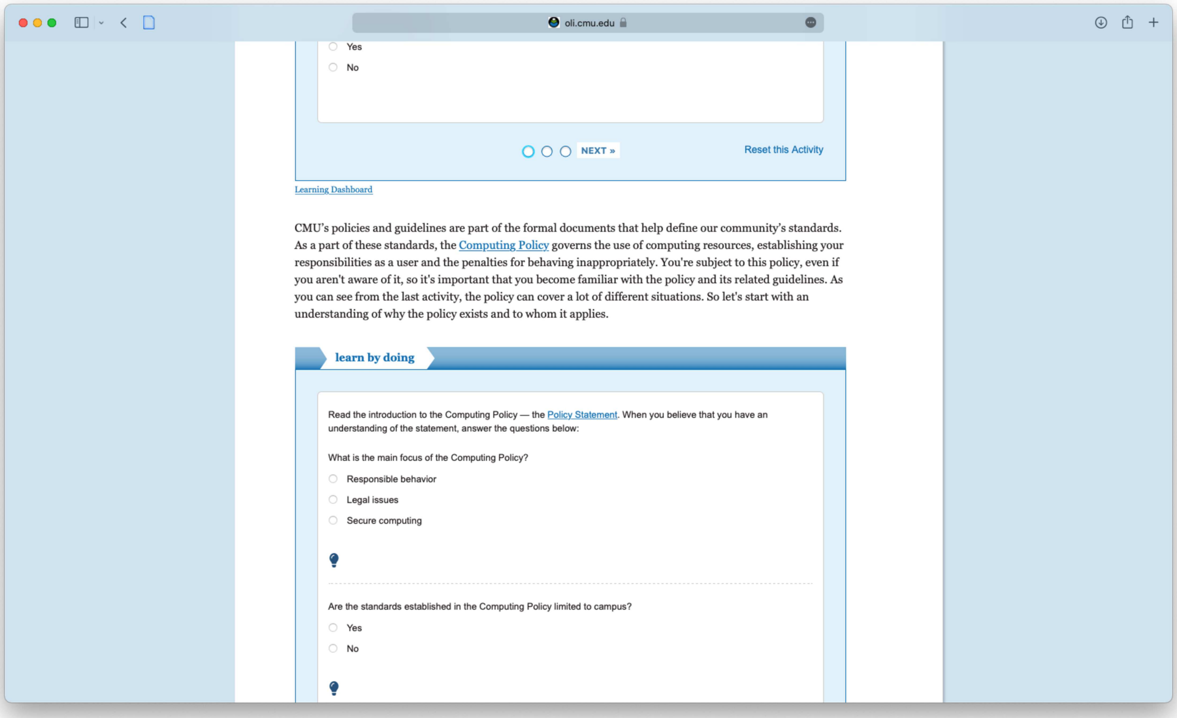
○ No

Downloaded five HD movies in one day.

○ Yes

○ No

Shared a song using BitTorrent.



○ Yes

○ No

○ ○ ○ NEXT »

Reset this Activity

Learning Dashboard

CMU's policies and guidelines are part of the formal documents that help define our community's standards. As a part of these standards, the Computing Policy governs the use of computing resources, establishing your responsibilities as a user and the penalties for behaving inappropriately. You're subject to this policy, even if you aren't aware of it, so it's important that you become familiar with the policy and its related guidelines. As you can see from the last activity, the policy can cover a lot of different situations. So let's start with an understanding of why the policy exists and to whom it applies.

learn by doing

Read the introduction to the Computing Policy — the Policy Statement. When you believe that you have an understanding of the statement, answer the questions below:

What is the main focus of the Computing Policy?

○ Responsible behavior

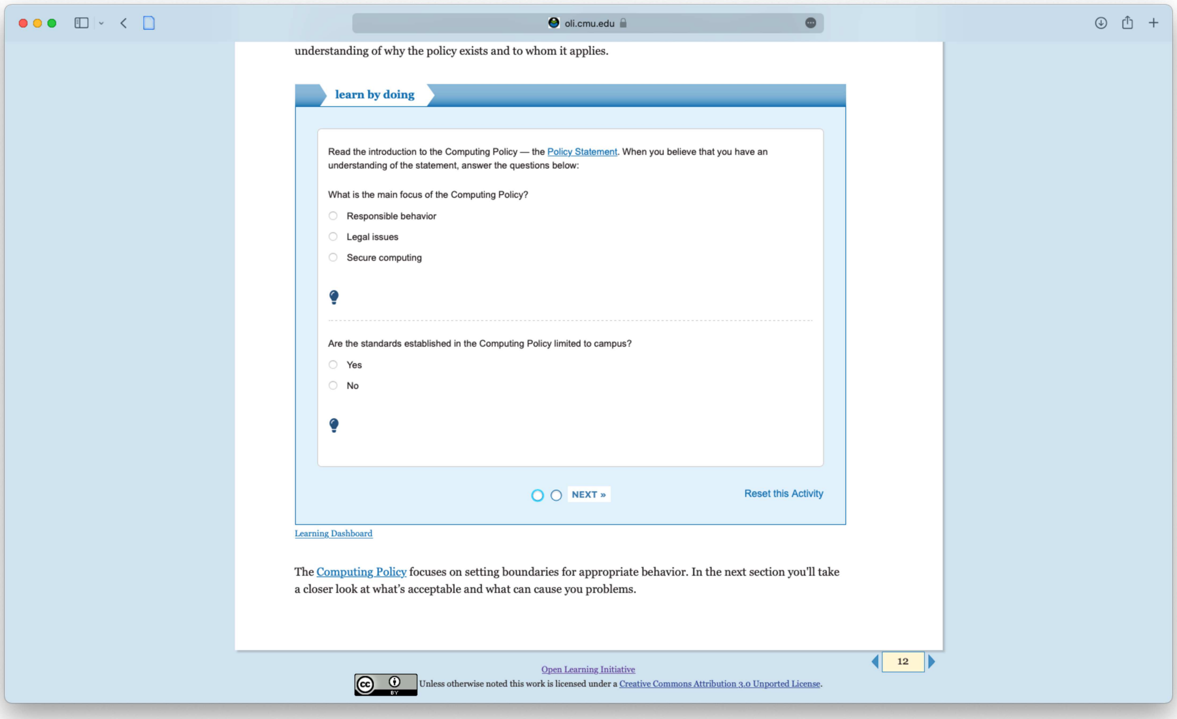
○ Legal issues

○ Secure computing

Are the standards established in the Computing Policy limited to campus?

○ Yes

○ No



understanding of why the policy exists and to whom it applies.

learn by doing

Read the introduction to the Computing Policy — the Policy Statement. When you believe that you have an understanding of the statement, answer the questions below:

What is the main focus of the Computing Policy?

○ Responsible behavior

○ Legal issues

○ Secure computing

Are the standards established in the Computing Policy limited to campus?

○ Yes

○ No

○ ○ NEXT »

Reset this Activity

Learning Dashboard

The Computing Policy focuses on setting boundaries for appropriate behavior. In the next section you'll take a closer look at what's acceptable and what can cause you problems.

Open Learning Initiative

12

Details of Sample

Table S1

Number of students who agreed to have their data analyzed in the Psychology Course of Study 1 divided by exam, pretest, and interaction data available.

	Did Not Complete Exam		Completed Exam			
Number of Quizzes Completed				Completed Pretest		**Grand Total**
	Did Not Complete Pretest	Completed Pretest	Did Not Complete Pretest	Interaction Data	Missing Interaction Data	
0	1,565					**1,565**
1	61	1,345	1	2		**1,409**
2	21	533		3	3	**560**
3	14	362		2	1	**379**
4	7	262	1	5		**275**
5	4	219		2		**225**
6	4	123		3	1	**131**
7		71	1	9		**81**
8		56		8		**64**
9	1	60	1	27	1	**90**
10	1	29	4	54	6	**94**
11	1	25	14	640	62	**742**
Grand Total	**1,679**	**3,085**	**22**	**755**	**74**	**5,615**

Table S2

Number of students who agreed to have their data analyzed in the Psychology Course of Study 1 divided by exam, pretest, and interaction data available.

			Completed Exam		
% Pretests Completed	**% Exams Completed**	**Did Not Complete Exam**	**Interaction Data**	**Missing Interaction Data**	**Grand Total**
0%	0%				
	25%			1	1
	50%			2	2
	75%			1	1
	100%	2		7	9
25%	0%	8			8
	25%	8	1		9
	50%	3		1	4
	75%		3	4	7
	100%	2	15	16	33
50%	0%			1	1
	25%	3		2	5
	50%	5		2	7
	75%		3	3	6
	100%	3	25	29	57
75%	0%				
	25%	1			1
	50%	1	1	2	4
	75%		3	4	7
	100%	1	40	46	87
100%	0%				
	25%			3	3
	50%		1		1
	75%	2	11	18	31
	100%	6	600	622	1,228
Grand Total		**45**	**703**	**764**	**1,512**

Measures

Table S3

Description of the measures, operationalization and how they were derived from student data.

Type Measure	Measure	Operationalization	Features extracted
Outcomes	Learning outcomes	Quiz grades for each module (Study 1 and Study 2) and for each learning objective (Study 2)	Average quiz grade for each module and student (Study 1 and Study 2) and across questions for each learning objectives (Study 2)
	Initial knowledge	Initial pretest for the entire course (Study 1) or for each module and/or learning objective (Study 2)	Average pretest grade (Study 1), average pretest quiz grade (Study 2), and average across questions in the pretest quiz for each learning objective (Study 2)
	Module difficulty	Average performance on the module activities	Average performance in all activities completed for student and module (Study 1 and Study 2)
Resource use	Doing	Percentage of activities completed in the module out of all available activities	For each student and module, count the number of activities the student interacted with and divide by the total number of activities in that module.
	Reading	Percentage of pages accessed in the module out of all available pages	For each student and module, count the number of pages the student opened and spent more than a quick amount of time on before completing an activity in the same page and divide by total

			number of pages in that module.
	Coverage	Percentage of learning objectives for which student completed activities out of all learning objectives with activities available.	Count the number of learning objectives with one or more activities completed for each module and student and divide by total number of available learning objectives in that module.
	Repetition	Number of activities the student completed only once (unique activities) or more than once (repeated activities)	Count the number of activities completed only once, and number of activities completed more than once.

Regression Results

Analyses of dosage effect (Study 1)

R code:

```
m1a <- lmer(zquizGrade ~ zpretestGrade+(1|ds_anon_user_id)+(1|module),data =
dbs[dbs$dataset=="ds863",])

m1b <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+(1|ds_anon_user_id)+(1|module),data =
dbs[dbs$dataset=="ds863",])

m1c <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+zpercentActivities+(1|ds_anon_user_id)+(1|module),data
= dbs[dbs$dataset=="ds863",])

m1d <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+zpercentActivities+zpercOutsideDo+(1|ds_anon_user_id)+
(1|module),data = dbs[dbs$dataset=="ds863",])

m1e <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+zpercentActivities+zpercOutsideDo+zpercentPages+(1|ds
_anon_user_id)+(1|module),data = dbs[dbs$dataset=="ds863",])

m1f <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+zpercentActivities+zpercOutsideDo+zpercentPages+zperc
OutsideRead+(1|ds_anon_user_id)+(1|module),data = dbs[dbs$dataset=="ds863",])

m1g <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+zpercOutsideDo+zpercentActivities*zpercentPages+zperc
OutsideRead+(1|ds_anon_user_id)+(1|module),data = dbs[dbs$dataset=="ds863",])
```

Table S4

Model comparison results for the analyses of dosage in Study 1

Model	Predictors	Comparison model	AIC	BIC	DF	χ^2	*p*
m1a	Pretest	-	16812	16846			
m1b	Pretest, **Correctness**	m1a	16676	16717	1	137.95	< .0001
m1c	Prestest, Correctness, **%Activity**	m1b	16545	16593	1	132.46	< .0001
m1d	Prestest, Correctness, %Activity, **%Activity_out**	m1c	16541	16595	1	6.56	0.010
m1e	Prestest, Correctness, %Activity, %Activity_out, **%Read**	m1d	16540	16601	1	2.84	0.091
m1f	Prestest, Correctness, %Activity, %Activity_out, %Read, **%Read_out**	m1e	16537	16605	1	5.09	0.024
m1g	Prestest, Correctness, %Activity, %Activity_out, %Read, %Read_out **%Activity *x* %Read**	m1f	16535	16610	1	3.25	0.071

Note. Bolded predictor is the added predictor in that given model.

Table S5

Regression results for the full model (m1g) for the dosage analyses Study 1

Effect	*Estimate*	*SE*	*95% CI*		*d*
			LL	*UL*	
Intercept	*-0.014*	*0.114*	*-0.247*	*-0.079*	
Pretest	*-0.119*	*0.021*	*-0.160*	*0.093*	*-0.440*
Correctness	*0.067*	*0.013*	*0.039*	*0.100*	*0.119*
% Activities completed outside the module	*0.060*	*0.021*	*0.019*	*0.195*	*0.148*
% Activities completed	*0.163*	*0.016*	*0.131*	*0.052*	*0.248*
% Pages read	*0.027*	*0.013*	*0.002*		*0.053*
% Pages read outside the module	*-0.045*	*0.021*	*-0.087*	*-0.003*	*-0.132*
*% Activities completed * % Pages read*	*-0.020*	*0.011*	*-0.043*	*0.002*	*-0.043*

Note. CI = confidence interval; LL = lower limit; UL = upper limit. All variables were z-scored.

Analyses of coverage (Study 1)

R code:

```
m2a <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentActivities+zpercentPages+(1|ds_anon_user_id)+(1|module),data=dbs_psych_with_kc)

m2b <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentActivities+zpercentKCs+zpercentPages+(1|ds_anon_user_id)+(1|module),data=dbs_psych_with_kc)

m2c <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentActivities*zpercentKCs+zpercentPages+(1|ds_anon_user_id)+(1|module),data=dbs_psych_with_kc)
```

Table S6

Model comparison results for the analyses of coverage in Study 1

Model	Predictors	Comparison model	AIC	BIC	DF	χ^2	*p*
m2a	Prestest, Correctness, %Activity, %Read	-	16544	16599			
m2b	Prestest, Correctness, %Activity, %Read, **%LOs**	m2a	16546	16608	1	0.36	.549
m2c	Prestest, Correctness, % Activity, %Read, %LOs, **%Activity *x* %LOs**	m2b	16539	16607	1	8.96	.003

Note. Bolded predictor is the added predictor in that given model.

Table S7
Regression results for the full model (m2c) for the analyses of coverage Study 1.

Effect	Estimate	SE	95% CI		d
			LL	UL	
Intercept	-0.083	0.116	-0.318	0.153	
Pretest	-0.122	0.021	-0.162	-0.081	-0.452
Correctness	0.082	0.015	0.053	0.111	0.137
% Activities completed	0.128	0.026	0.076	0.180	0.118
% LOs completed	0.124	0.043	0.039	0.209	0.069
% Pages read	0.020	0.012	-0.004	0.045	0.041
*% Activities * % LOs*	0.075	0.025	0.026	0.124	0.073

Note. CI = confidence interval; LL = lower limit; UL = upper limit. All variables were z-scored.

Analyses of repetition (Study 1)

R code:

```
m3a  <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentKCs+zpercentPages+(1|ds_anon_user_id)+(1|module),data=dbs_psych_with_kc[!is.na(dbs_psych_with_kc$znActivities_repeated),])

m3b  <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentKCs+zpercentPages+znActivities_unique+(1|ds_anon_user_id)+(1|module),data=dbs_psych_with_kc[!is.na(dbs_psych_with_kc$znActivities_repeated),])

m3c  <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentKCs+zpercentPages+znActivities_unique+znActivities_repeated+(1|ds_anon_user_id)+(1|module),data=dbs_psych_with_kc[!is.na(dbs_psych_with_kc$znActivities_repeated),])

m3d  <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentKCs+zpercentPages+znActivities_unique*znActivities_repeated+(1|ds_anon_user_id)+(1|module),data=dbs_psych_with_kc[!is.na(dbs_psych_with_kc$znActivities_repeated),])
```

Table S8

Model comparison results for the analyses of repetition in Study 1

Model	Predictors	Comparison model	AIC	BIC	DF	χ^2	*p*
m3a	Pretest, Correctness, %LO, %Read	-	15458	15512			
m3b	Pretest, Correctness, %LO, %Read, **#Unique Activities**	m3a	15454	15515	1	5.69	0.017
m3c	Pretest, Correctness, %LO, %Read, #Unique Activities, **#Repeat Activities**	m3b	15453	15520	1	3.51	0.061
m3d	Pretest, Correctness, %LO, %Activities, %Read, #Unique Activities, #Repeat Activities, **#Unique *x* #Repeated**	m3c	15454	15529	1	0.19	0.659

Note. Bolded predictor is the added predictor in that given model.

Table S9

Regression results for the full model (m3d) for the analyses of repetition for Study 1.

Effect	Estimate	SE	95% CI		d
			LL	UL	
Intercept	-0.079	0.114	-0.313	0.155	
Pretest	-0.122	0.021	-0.164	-0.080	-0.44
Correctness	0.092	0.015	0.063	0.122	0.16
% LOs Completed	0.000	0.061	-0.121	0.120	0.00
% Pages Completed	0.025	0.013	-0.001	0.051	0.05
# Unique Activities Completed	0.150	0.062	0.029	0.271	0.06
# Repeated Activities Completed	-0.019	0.013	-0.045	0.007	-0.04
*# Unique Activities Completed * # Repeated Activities Completed*	-0.006	0.014	-0.035	0.022	-0.01

Note. CI = confidence interval; LL = lower limit; UL = upper limit. All variables were z-scored.

Analyses of dosage effect (Study 2)

R code:

```
m4a <- lmer(zquizGrade ~ zpretestGrade+(1|ds_anon_user_id)+(1|module),data =
dbs[!dbs$dataset=="ds863",])

m4b <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+(1|ds_anon_user_id)+(1|module),data =
dbs[!dbs$dataset=="ds863",])

m4c <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+zpercentActivities+(1|ds_anon_user_id)+(1|module),data
= dbs[!dbs$dataset=="ds863",])

m4d <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+zpercentActivities+zpercOutsideDo+(1|ds_anon_user_id)+
(1|module),data = dbs[!dbs$dataset=="ds863",])

m4e <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+zpercentActivities+zpercOutsideDo+zpercentPages+(1|ds
_anon_user_id)+(1|module),data = dbs[!dbs$dataset=="ds863",])

m4f <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+zpercentActivities+zpercOutsideDo+zpercentPages+zperc
OutsideRead+(1|ds_anon_user_id)+(1|module),data = dbs[!dbs$dataset=="ds863",])

m4g <- lmer(zquizGrade ~
zpretestGrade+zcorrectness_alltries+zpercOutsideDo+zpercentActivities*zpercentPages+zperc
OutsideRead+(1|ds_anon_user_id)+(1|module),data = dbs[!dbs$dataset=="ds863",])
```

Table S10
Model comparison results for the analyses of dosage in Study 2.

Model	Predictors	Comparison model	AIC	BIC	DF	χ^2	*p*
m4a	Pretest		5209	5236			
m4b	Pretest, **Correctness**	M4a	5175	5209	1	35.16	<.0001
m4c	Prestest, Correctness, **%Activity**	M4b	5153	5192	1	24.39	<.0001
m4d	Prestest, Correctness, %Activity, **%Activity_out**	M4c	5154	5198	1	1.34	.247
m4e	Prestest, Correctness, %Activity, %Activity_out, **%Read**	M4d	5155	5206	1	0.32	.571
m4f	Prestest, Correctness, %Activity, %Activity_out, %Read, **%Read_out**	M4e	5155	5211	1	1.88	.171
m4g	Prestest, Correctness, %Activity, %Activity_out, %Read, %Read_out, **%Activity x %Read**	M4g	5137	5198	1	20.89	<.0001

Note. Bolded predictor is the added predictor in that given model.

Table S11

Regression Results for the full model (m4g) for the analyses of dosage in Study 2

Effect	*Estimate*	*SE*	*95% CI*		*d*
			LL	*UL*	
Intercept	*0.022*	*0.209*	*-0.438*	*0.482*	
Pretest	*0.061*	*0.022*	*0.018*	*0.103*	*0.131*
Correctness	*0.045*	*0.025*	*-0.003*	*0.093*	*0.084*
% Activities completed outside the module	*0.027*	*0.027*	*-0.026*	*0.081*	*0.051*
% Activities completed	*0.217*	*0.032*	*0.155*	*0.279*	*0.311*
% Pages read	*-0.018*	*0.022*	*-0.061*	*0.026*	*-0.036*
% Pages read outside the module	*-0.036*	*0.027*	*-0.089*	*0.016*	*-0.074*
*% Activities completed * % Pages read*	*-0.042*	*0.009*	*-0.059*	*-0.024*	*-0.208*

Note. CI = confidence interval; LL = lower limit; UL = upper limit. All variables were z-scored.

Analyses of coverage (Study 2)

R code:

```
m5a <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentActivities+zpercentPages+(1|ds_anon_user_id)+(1|module),data=dbs_computing_with_kc)

m5b <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentActivities+zpercentKCs+zpercentPages+(1|ds_anon_user_id)+(1|module),data=dbs_computing_with_kc)

m5c <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentActivities*zpercentKCs+zpercentPages+(1|ds_anon_user_id)+(1|module),data=dbs_computing_with_kc)
```

Table S12

Model comparison results for the analyses of coverage in Study 2.

Model	Predictors	Comparison model	AIC	BIC	DF	χ^2	p
m5a	Prestest, Correctness, %Activity, %Read	-	5154.6	5199.2			
m5b	Prestest, Correctness, %Activity, %Read, **%LOs**	m5a	5152.6	5202.8	1	3.91	.048
m5c	Prestest, Correctness, %Activity, %Read, %LOs, **% Activity *x* %LOs**	m5b	5133.4	5189.2	1	21.25	<.001

Note. Bolded predictor is the added predictor in that given model.

Table S13

Regression results for the full model (m5c) for the analyses of coverage Study 2.

Effect	*Estimate*	*SE*	*95% CI*		*d*
			LL	*UL*	
Intercept	*0.036*	*0.212*	*-0.429*	*0.502*	
Pretest	*0.064*	*0.022*	*0.021*	*0.106*	*0.138*
Correctness	*0.037*	*0.025*	*-0.012*	*0.086*	*0.067*
% Activities completed	*0.224*	*0.039*	*0.146*	*0.300*	*0.258*
% LOs completed	*0.025*	*0.028*	*-0.030*	*0.080*	*0.041*
% Pages read	*-0.015*	*0.024*	*-0.062*	*0.031*	*-0.029*
*% Activities * % LOs*	*-0.046*	*0.010*	*-0.065*	*-0.026*	*-0.210*

Note. CI = confidence interval; LL = lower limit; UL = upper limit. All variables were z-scored.

Analyses of repetition (Study 2; module level)

R code:

```
m6a <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentKCs+zpercentPages+(1|ds_anon_user_id)+(1|module),data=dbs_computing_with_kc[!is.na(dbs_computing_with_kc$znActivities_repeated),])
```

```
m6b <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentKCs+zpercentPages+znActivities_unique+(1|ds_anon_user_id)+(1|module),data=dbs_computing_with_kc[!is.na(dbs_computing_with_kc$znActivities_repeated),])
```

```
m6c <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentKCs+zpercentPages+znActivities_unique+znActivities_repeated+(1|ds_anon_user_id)+(1|module),data=dbs_computing_with_kc[!is.na(dbs_computing_with_kc$znActivities_repeated),])
```

```
m6d <-
lmer(zquizGrade~zpretestGrade+zcorrectness_alltries+zpercentKCs+zpercentPages+znActivities_unique*znActivities_repeated+(1|ds_anon_user_id)+(1|module),data=dbs_computing_with_kc[!is.na(dbs_computing_with_kc$znActivities_repeated),])
```

Table S14

Model comparison results for the analyses of repetition using module-level data in Study 2.

Model	Predictors	Comparison model	AIC	BIC	DF	χ^2	*p*
m6a	Pretest, Correctness, %LO, %Read	-	5162.0	5206.7			
m6b	Pretest, Correctness, %LO, %Read, **#Unique Activities**	m6a	5138.0	5188.2	1	26.03	<.001
m6c	Pretest, Correctness, %LO, %Read, #Unique Activities, **# Repeat Activities**	m6b	5136.6	5192.4	1	3.41	.065
m6d	Pretest, Correctness, %Activities, %Read, #Unique Activities, #Repeat Activities, **#Unique *x* #Repeat**	m6c	5133.0	5194.3	1	5.63	.018

Note. Bolded predictor is the added predictor in that given model.

Table S15

Regression results for the full model (m6d) for the analyses of repetition using module-level data for Study 2.

Effect	Estimate	SE	95% CI		d
			LL	UL	
Intercept	0.011	0.204	-0.436	0.459	
Pretest	0.065	0.022	0.022	0.107	0.140
Correctness	0.047	0.024	0.000	0.094	0.090
% LOs completed	-0.078	0.043	-0.163	0.007	-0.083
% Pages read	-0.030	0.023	-0.076	0.016	-0.059
# Unique activities completed	0.256	0.047	0.165	0.348	0.250
# Repeated activities completed	0.030	0.037	-0.041	0.102	0.038
# Unique activities* # Repeated activities	-0.027	0.011	-0.049	-0.005	-0.108

Note. CI = confidence interval; LL = lower limit; UL = upper limit. All variables were z-scored.

Analyses of repetition (Study 2; Learning Objective level)

Table S16

Model comparison results for the analyses of repetition using objective level data in Study 2.

Model	Predictors	Comparison model	AIC	BIC	DF	χ^2	*p*
m7a	Pretest, Correctness,	-	7204.3	7240.6			
m7b	Pretest, Correctness, **#Unique Activities**	m7a	7195.1	7237.4	1	11.25	<.001
m7c	Pretest, Correctness, #Unique Activities, **#Repeat Activities**	m7b	7197.1	7245.5	1	0.01	0.913
m7d	Pretest, Correctness, #Unique Activities, #Repeat Activities, **#Unique *x* #Repeat**	m7c	7196.7	7251.2	1	2.34	0.126

Note. Bolded predictor is the added predictor in that given model.

Table S17
Regression results for the full model (m7d) for the analyses of repetition using objective-level data for Study 2

Effect	Estimate	SE	95% CI		d
			LL	UL	
Intercept	-0.028	0.201	-0.438	0.381	
Pretest	0.056	0.015	0.027	0.086	0.134
Correctness	0.048	0.016	0.016	0.079	0.108
# Unique activities completed	0.063	0.019	0.026	0.100	0.127
# Repeated activities completed	-0.008	0.015	-0.037	0.021	-0.020
# Unique activities* # Repeated activities	-0.029	0.019	-0.065	0.008	-0.055

Note. CI = confidence interval; LL = lower limit; UL = upper limit. All variables were z-scored.