Is the Doer Effect a Causal Relationship? How Can WE Tell and Why It's Important

Kenneth R. Koedinger
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15201
koedinger@cmu.edu

Elizabeth A. McLaughlin
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15201
mimim@cs.cmu.edu

Julianna Zhuxin Jia
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15201
zhuxinj@andrew.cmu.edu

Norman L. Bier
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15201
nbier@cmu.edu

ABSTRACT
The "doer effect" is an association between the number of online interactive practice activities students' do and their learning outcomes that is not only statistically reliable but has much higher positive effects than other learning resources, such as watching videos or reading text. Such an association suggests a causal interpretation--more doing yields better learning--which requires randomized experimentation to most rigorously confirm. But such experiments are expensive, and any single experiment in a particular course context does not provide rigorous evidence that the causal link will generalize to other course content. We suggest that analytics of increasingly available online learning data sets can complement experimental efforts by facilitating more widespread evaluation of the generalizability of claims about what learning methods produce better student learning outcomes. We illustrate with analytics that narrow in on a causal interpretation of the doer effect by showing that doing within a course unit predicts learning of that unit content more than doing in units before or after. We also provide generalizability evidence across *four different courses involving over 12,500 students* that the learning effect of doing is about *six times greater* than that of reading.

CCS Concepts
• **Applied computing** → **E-learning**

• **Applied computing** → **Computer- managed instruction**

Keywords
Learn by doing; prediction; course effectiveness; doer effect; learning engineering

LAK '16, April 25-29, 2016, Edinburgh, United Kingdom
 ISBN 978-1-4503-4190-5/16/04...$15.00
DOI: http://dx.doi.org/10.1145/2883851.2883957

1. INTRODUCTION
One general challenge for learning analytics in particular, and for learning science and practice in general, is how to reliably determine what are the most effective methods for supporting learning and under what circumstances do those methods work. We argue that learning analytics has something distinct and important to offer as an answer to these questions. Whereas random assignment controlled experimentation is considered the gold standard for determining whether a method of learning is effective [9], it does not provide evidence on whether that learning method will generalize to other courses and course contexts. In contrast, the increasing availability of process and outcome data from online courses [16] makes it possible to investigate the generalizability of associations between learning method and outcomes. Because such data comes from naturally occurring variations in use rather than from random assignment, we cannot be sure that those associations are causal. However, such data adds evidence for generalization (or lack thereof) that comes at a much lower monetary and social/ethical cost than would be needed to do random assignment experiments across all of these naturally occurring contexts [cf., 3].

In more technical terms, an experiment provides strong *internal validity* for causal inference but provides no *external validity* for generalization of the method to contexts not sampled in that experiment. Analysis of associations of method and outcome threatens *internal validity,* but doing so across many naturally occurring contexts provides *external validity* at a much lower cost than doing experiments in all these contexts.

Without the costs of designing and executing any experiments, we have five data sets that were collected as a natural part of five courses from four different content areas. While most experiments typically evaluate one method (with two conditions), these data sets allow us to analyze outcomes associated with three different methods (doing activities, reading text, or watching video lectures) for one course and two different methods (doing and reading) for four courses. To do 11 controlled experiments in these real world contexts would be a much more costly undertaking.

Another key point of the current paper is to explore analytic techniques that help eliminate alternative causal interpretations so as to get closer to causal inference even when the data is correlational in character. In particular, we explore the use of intermediate course unit quiz data to evaluate whether the same

student reveals variation in choices across units that is associated with better learning outcomes. This cross-unit analysis is also another test of generalization of learning method effects across the different content types and goals of each unit.

Lower Cost Analytics Enhances Generalizability

In a randomized controlled experiment of the effectiveness of a method for supporting learning, students are randomly assigned to either a treatment condition, where the method is used, or a control condition, where it is not. The effect of the conditions are compared on how well students do on some common outcome measure of learning achievement. Many random assignment experiments on learning have been run in labs [e.g., 11]. To use an example relevant to this paper, some experiments have compared whether having students practice retrieving facts (e.g., What is the Chinese word for teacher?) yields better learning than having students study facts (e.g., The Chinese word for teacher is lao3shi1). Such experiments have demonstrated, with statistical reliability, better long term learning outcomes from retrieval practice (also called "testing") than from studying [12, 14]. Random assignment experiments like these in the lab have high *internal validity* [1, 18] meaning that we can be confident that the method (e.g., more "testing") is causing the better outcomes.

Some critics wonder whether these lab results generalize to the classroom and thus randomized experiments run within real courses are sometimes performed, though often at greater cost in terms of both real and social/ethical capital. Classroom studies of the testing effect [e.g., 8] add ecological validity by being performed with real students and in the context of real courses. Even here the few experiments in a limited number of course contexts may leave a critic wondering whether these testing effects generalize across all course content or might be limited to certain kinds of content or contexts. In fact, many of these studies have focussed on the learning of facts and verbally communicated content [6, 11]. Perhaps the testing effect is less relevant to the learning of skills or principles. We do not have sufficient external validity evidence. Research on worked examples [e.g., 4, 11, 13, 15] suggests limits on the testing effect -- these studies find that too much practice and not enough studying can yield poorer learning outcomes. More generally, the KLI Framework [6] outlines empirical and theoretical reasons to believe that many methods for learning and instruction do not generalize across course content, that is, they work well for some kinds of knowledge acquisition but not for others (e.g., sense making support is best for learning principles but less for skills and arbitrary facts, inductive learning support is best for skills but not for facts, and memory supports are best for facts).

In addition to evidence for the testing effect, there is broader advocacy and evidence for related notions of learning by doing [e.g., 2] and active learning [e.g., 20]. These terms are arguably less precise than the testing effect as they cover a wider set of approaches and are more loosely defined. Testing effect studies typically involve immediate feedback on responses that provide students with an example correct response if they fail to retrieve one themselves. In cases where retrieval practice conditions do not involve feedback, smaller learning outcomes are observed (e.g., see Table 1 in [14]). Timely feedback is employed in many learning by doing and active learning applications, however, learning by doing is also often used to refer to project-based, constructivist, or more open-ended inquiry approaches where instructional supports such as immediate interactive feedback on student progress are deemphasized and even discouraged [c.f., 4]. There is some large-scale evidence of classroom-based benefits of interventions emphasizing learning-by-doing or active learning including quasi- experiments with non-randomly assigned controls [e.g., 5, 20, 21] and classroom-based experimental trials with random assignment [e.g., 10,17]. However, in all these cases, the treatment conditions vary in many ways from the control conditions so as to not isolate active doing as a causal ingredient.

It should be clear that determining causal relationships is important for scientific and practical reasons because causal relationships provide a path toward explanatory theory and a path toward reliable and replicable practical application. Further, we need evidence for causes that generalize across a wide variety instructional contexts and course content. If we can be increasingly certain a learning method is causally related to more optimal learning across a wide variety of contexts and content, then that method should be used to guide course design and students should be encouraged to use it. Coupling evidence from both experiments and analysis of naturally occurring high-volume data appears an effective way to increase generalizable certainty.

Toward Explaining the "Doer Effect" and Exploring its Generality

In prior work, we found that different student choices of learning methods (e.g., doing interactive activities, reading online text, or watching online lecture videos) are associated with learning outcomes [7]. More usage in general is associated with higher outcomes, but especially for doing activities which has an estimated 6x greater impact on total quiz and final exam performance than reading or video watching. One open question regarding this "doer effect" is whether the observed association is indicative of a causal relationship, that is, that students learn more as a consequence of doing more. Alternatively, there may be no causal link between the two, but rather some "third variable" common cause, such as general student motivation to learn, that leads to both more doing and better learning. The fact that the effect of doing activities is much stronger than that of watching videos or reading text suggests that the third variable cannot be general to all learning methods but would have to be particular to doing -- something like: students who are generally good learners desire to demonstrate their competence ("show off") and doing is a better way to do so than reading or watching.

In this paper we introduce an analysis approach designed to probe any such general student trait explanation in contrast to the causal explanation. This technique relies on course data involving repeated unit assessments throughout the course with process data on student use of different learning methods relevant to the unit between these assessments. If the causal explanation is correct, then the amount of doing a student chooses to engage in during a unit should be predictive of their performance on that unit assessment above beyond any effect of the amount of doing outside that unit. In contrast, if some general trait is an explanation then the amount of doing outside a unit should be equally predictive (or more because there is more data outside a unit) of that unit's assessment results as the amount of doing within that unit. Statistically, a regression model should reveal no within-unit effect above and beyond the outside-unit effect.

A second open question we explore is whether the doer effect generalizes across multiple online courses. We do so with data from four Open Learning Initiative online courses (Concepts in Computing, Introduction to Biology, Introduction to Psychology, and Statistical Reasoning) on associations between student variation in the amount of doing and their learning outcomes and between student variation in the amount of reading and their

learning outcomes. We focus our investigations on these specific research questions:

1. Can we use cross-course performance data to narrow down the possible causal interpretations of the doer effect?
    a. How do individual student resource choices vary across units of a course?
    b. Is student performance in each unit better predicted by how much they do in those units than by how much they do in other units?
    c. Might course unit prerequisite relationships contribute to cross-unit doer effects?
2. Does the doer effect generalize to different courses or to different students using similar course materials?

2. CROSS-COURSE PERFORMANCE DATA TOWARD BETTER CAUSAL EVIDENCE FOR THE DOER EFFECT

2.1 Method: Context and Nature of Data

This analysis involves data from students taking the *Introduction to Psychology as a Science* MOOC course offered by Georgia Institute of Technology through Coursera. This is the same data used in [7] and more details about the course, about student characteristics, and about factors leading to drop out can be found in that paper. Here we focus on the students who finished the course and took all or most of the 11 quizzes (N=1154). Of these students, most of them (N=1051) opt to make at least some use of corresponding online materials (readings and interactive activities) from the Open Learning Initiative (OLI) course titled *Introduction to Psychology* offered by Carnegie Mellon University.

The course is designed to introduce college students to the broad topics in the discipline of psychology. The 12-week course includes video lectures on each topic (e.g., biopsychology, sensation and perception, learning) presented by the course professor. In addition, each topic is aligned with modules from the *Introduction to Psychology* OLI course that students are encouraged to use as an online textbook and practice environment. The course syllabus maps the topics of the lectures to the OLI modules for each week. Thus, if students take advantage of the course offerings their learning environment includes watching videos in Coursera, reading OLI text pages, and doing OLI interactive activities. The Coursera portion also includes a discussion board, which we do not analyze here, but it is addressed in other research [19].

Interactive activities are aligned with course learning objectives and are embedded in the course content. They provide opportunities for students to test their understanding of concepts and to practice skills. Such learning opportunities take various formats (e.g., multiple choice questions, interactive simulations, drop and drag, matching, and other options) and deliver immediate tailored feedback as-needed (e.g., when a selected answer is incorrect) or as-requested (e.g., in the form of a hint). Many activities are multiple-choice questions, but others, like those shown in Figure 1, provide other forms of response selection (1a) and response construction, including the open-ended submit and compare (1b). In all cases, students have immediate access to correct responses.

After each video lecture for the first 11 weeks of the course, there was a quiz. After the final week, students took a cumulative final exam. For our by-unit analysis of student activity to outcome associations, we focus on their activities in the 11 units associated with the 11 weekly quizzes. Each quiz had 10 items on it. Across all students and all quizzes the average quiz performance was 8.2 out of 10.

Although the course had weekly quizzes aligned to the content for that week, students had some flexibility for when they took a specific week's quiz. Given this autonomy and the resulting variance in quiz start times, resource data frequencies were individualized per student. Two factors were used for attaching unit resource data to a weekly quiz: time and content. Data was mapped to course content using the syllabus and OLI modules for reading and doing, and urls were used for watching (videos had an assigned unit number). For each student, all relevant data per quiz was tallied until the start time for that quiz, all remaining data was considered irrelevant. Therefore, all readings, activities and videos associated with the content before a quiz is taken are deemed relevant.

2.2 Results: Variation in Student Choices

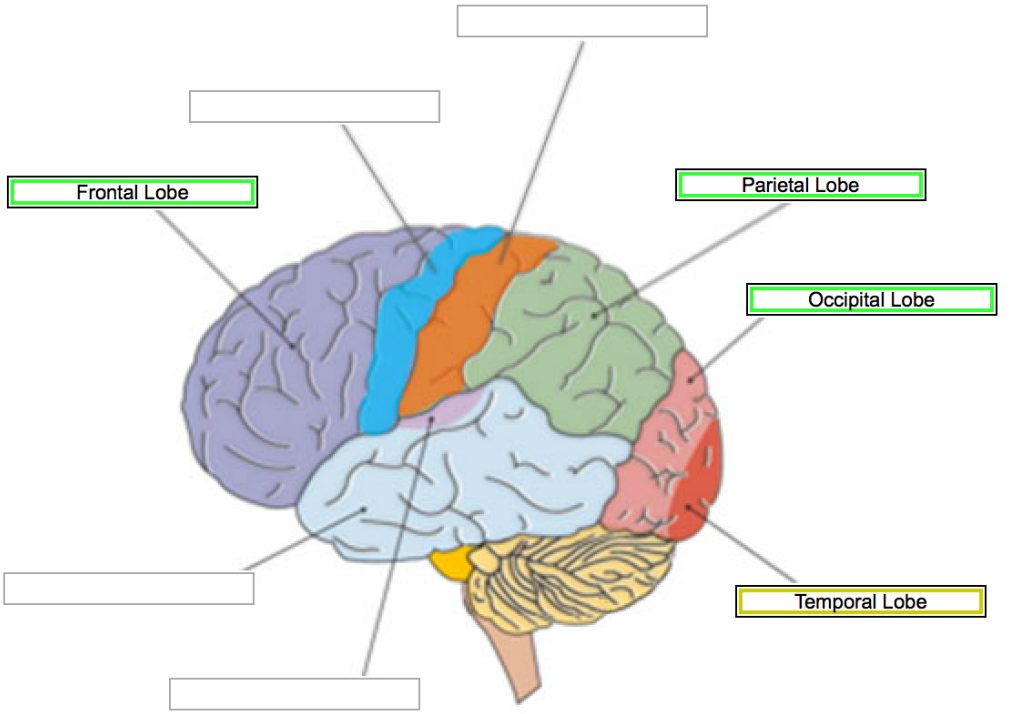
Before investigating whether within-unit choices better predict unit outcomes than outside-unit activities, we first verify that there is sufficient variation in individual student resource use choices across units of the course to justify our further analysis. We wanted to determine whether students vary in how active they are during each weekly unit of the course. If students who do a lot of activities always do a lot and those that do few always do few, then the by-unit analysis we propose will be uninformative.

To check for variability, we used the activity data from the 1051 students who accessed at least some pages or activities in OLI (103 students in our sample did not). Each of these students worked through 11 units, producing 11,561 (= 1051 x 11) student-unit combinations. For each of these student-unit combinations we compared students' level of activity within the target unit to their activity outside of it. No surprise, these measures are highly correlated, R = .68. However, there is variation. To investigate how much, for each quiz we grouped students into 5 groups (quintiles) based on their within-unit activity and 5 quintile groups based on their outside-unit activity. As shown in the bottom row of Table 1, outside-unit quintile boundaries produce reasonably consistently sized groups (a consequence of having lots of opportunity for different levels of outside-unit activity counts from the 10 units outside each unit). About 20% of students are in each of the lowest or 1st quintile (below about 8 outside-activities), the 2nd quintile (below about 185), 3rd quintile (below about 426), 4th quintile (below about 538), and the highest or 5th quintile (at or above about 538). Within-unit quintile boundaries vary more because the number of activities available and done within a unit changes quite a bit and in some cases is small enough to yield issues where whole number quintile cut-offs produce quintile groups of different sizes (e.g., in units 9-11 where there is a median of 15, 11, and 2 activities done, more than 40% of the students did 0 activities so there is no way to differentiate the first and second quintile -- all such students are in the 1st quintile and no students are in the second).

In 6266 instances or 54% of the student-unit combinations, the within-unit activity quintile was different from the outside-unit activity. In 1464 instances or 13% of the cases, the quintile was different by two levels (a difference of more between 20-40 percentile points). For example, of all the student-units in the 3rd

(a)

Complete the following diagram by dragging the labels at the bottom into the appropriate spots on the diagram.



That's incorrect. The temporal lobe is one of the four major sections of each hemisphere of the cerebral cortex.

(b)

Imagine that you are a brain scientist. In the following scenarios, select the best method to learn more about the person's brain and the presenting problem.

Hint

Laura, a patient, complains of debilitating migraine headaches. She has tried a number of medications, but nothing has relieved her symptoms. As a scientist, you propose that she try a new brain technique that might relieve her painful headaches. Name the brain technique and explain how it might relieve her symptoms.

Submit and Compare

Figure 1. A sample of interactive activities from the Brain Regions module of the OLI course used in the Psychology MOOC. The example on top (a) illustrates a machine-gradable alternative to multiple-choice with vastly more choices (8! = 40,320). The bottom of the figure (b) is an example of an open-ended "submit and compare" question, where students can compare their submitted response to an example correct response. In all OLI activities, students have immediate access to correct responses.

Table 1. Student activity within each unit compared with their activity outside that unit.

	Outside-Unit Activity Quintile					
Within-Unit Activity Quintile	Low 20%	2nd 20%	3rd 20%	4th 20%	High 20%	*Within Totals*
Lowest 20%	2244	1406	473	82	26	4231
2nd 20%	57	308	135	32	11	543
3rd 20%	53	505	1176	880	598	3212
4th 20%	5	34	401	989	1080	2509
Highest 20%	1	22	127	338	578	1066
Outside Totals	2360	2275	2312	2321	2293	11561

quintile of outside-unit activity (2312 of them), there are 473 cases (20%) where the within-unit activity is quite a bit lower (20- 40 percentile points) than outside-unit activity and 127 cases (5%) where the within-unit activity is quite a bit higher than outside activity. In summary, we do find lots of cases where students chose to do many fewer or many more activities than they tend to do otherwise. This natural variability opens the door to analyzing whether within-unit activity is predictive of learning unit content above and beyond outside-unit activity.

2.3 Results: Association of Within-Unit and Outside-Unit Choices with Learning Outcomes

To investigate the association of within-unit and outside-unit choices with learning outcomes, we used mixed effect linear regression modeling, implemented using lmer function in R, an open statistical application. We aggregated log data from Coursera and OLI into a file with 11561 rows for the 1051 students and each of the 11 units. The outcome or dependent measure is the unit quiz score for the given student and unit. To derive predictor (or independent) measures we developed an analytic script to extract from the log data the number of activities started, pages accessed, and video started within each unit. (Note: Doing so was no small effort, motivating a need for analytic script sharing that learnsphere.org is being designed to support.) These resource use counts were constrained to both be resources within the course content associated to that unit (i.e., a Coursera video within this unit's section of the syllabus or an OLI page or activity within an associated OLI unit) *and* used before the student took the associated quiz. For example, a resource done in week 1 but associated with unit 2 that is done before the quiz (even if in week 1) gets counted toward unit 2, but that same resource done after the unit quiz 2 is not counted. All student resource use that is not counted as within-unit by the above criteria is than counted as outside-unit (e.g., for unit 2 any resource associated with a different unit and any unit 2 resource used after the unit 2 quiz). Thus, for each student-unit row, we had within-unit and outside- unit counts for each of doing, reading, and watching. As in [7], we adjusted each student's reading score to only count pages accessed beyond the estimated minimum needed to access the number of activities that student did. We also converted all measures to Z scores (standard deviations from the mean) to aid interpretation, namely, to facilitate direct comparison of model parameter estimates.

Table 2. Within-unit and outside-unit effects of resource use on unit quiz performance.

Learning method	Location	Parameter Estimate	Std. Error	DF	t value	Pr(>|t|)
	(Intercept)	-0.015	0.068	12	-0.218	0.8312
Doing	*Within-unit*	0.195	0.011	9969	17.475	< 0.000001 ***
	Outside-unit	0.196	0.022	1389	8.736	< 0.000001***
Reading	*Within-unit*	0.015	0.009	10226	1.676	0.0937 .
	Outside-unit	-0.006	0.021	1184	-0.283	0.7770
Watching	*Within-unit*	0.036	0.009	10244	4.174	<0.00003 ***
	Outside-unit	-0.002	0.020	1215	-0.103	0.9182

Shown below is the R formula we used for this analysis, indicating both the statistical method, a linear mixed effect regression (lmer), and the regression formula (variables with "NR", for non-relevant, indicate the outside-unit counts):

```
lmer(Z.quiz.correct ~ (1|user.name) + (1|quiz.num) +
     Z.Activities + Z.NR.Activities +
     Z.Readings + Z.NR.Readings +
     Z.Video + Z.NR.Videos, data = b_a)
```

To adjust for general differences in student performance and unit quiz difficulty we included random effects in the model for both student and unit (coded as quiz.num). We report on analysis for the subset of registered OLI students (N = 939), since only OLI students have the option of doing and reading. All the significant results remain the same when we include all students.

The key findings are shown in Table 2. There are significant effects of within-unit and outside-unit doing, and within-unit video watching. Within-unit reading is marginal. Outside-unit reading and outside-unit watching are not significant.

We find that within-unit doing remains a large and higher significant predictor even after controlling for non-relevant choices. This result is consistent with a causal interpretation. Inspecting the parameter values we see, as before, a much larger association of doing with outcomes than watching or reading with outcomes. Whereas the prior whole course analysis [7] found about a 6 times greater effect of doing on outcomes than reading or watching, here we find a 13 times bigger effect of doing than reading and a greater than 5 times effect of doing than video watching.

We also see, at least for doing, a significant effect of outside-unit resourse use. This result may indicate some third variable yielding both higher general doing and better outcomes. One possibility is that this third variable is indeed some general student trait -- a third variable account for the causing better learning, is that there are prerequisite. activity effect. Another possibility, consistent with doing causing better learning, is that there are prerequisite relationships between units such that doing more activities in an earlier unit, say unit 4, not only improves learning of that content but better prepares the student for better learning from a related subsequent unit, say unit 6.

2.4 Results: Might Prerequisites Account for the Doer Effect for Outside-Unit Doing

To test for the possibility that the large outside-unit effect of doing may be a consequence of better learning of prerequisites, we split the outside-unit counts into before-unit and after-unit counts. We ran a related R analysis as shown below (as above, the model normalizes for general student competence and for quiz difficulty by including random effects for these):

```
lmer (Z. quiz.correct ~ (1|user.name) + (1|quiz.num) +
     Z.BF.act + Z.activities + Z.AF.act +
     Z.BF.reading + Z.reading + Z.AF.reading +
     Z.BF.Video + Z.Video + Z.AF.Video, data = b_a2)
```

We report on results considering only units 2 to 10 in this before and after analysis, since resources are not available before unit 1 and the resources used after unit 11 are of a different character (most being accessed to study for the final exam). As above, we limited to only registered OLI students. All the significant results remain the same when we include either or both all students and/or all units.

We find that the doer effect is now strongest for doing *within* the target unit as compared to doing *before* or *after*. This stronger effect for doing within than before or after comes

Table 3. Before within, after unit effects of resource use on unit quiz performance.

Learning method	Location	Normalized Estimate	Std. Error	df	t value	Pr(>|t|)
	(Intercept)	0.059	0.049	11	1.344	0.20506
Doing	*Before*	0.143	0.016	1861	8.840	< 0.00 ***
	Within	0.181	0.011	8892	16.973	< 0.00 ***
	After	0.078	0.014	3063	5.402	0.00 ***
Reading	*Before*	0.008	0.015	1635	0.579	0.56278
	Within	0.010	0.008	9233	1.224	0.22116
	After	-0.013	0.012	2603	-1.028	0.30410
Watching	*Before*	0.054	0.014	2432	3.853	0.00012 ***
	Within	0.025	0.008	9463	3.123	0.00180 **
	After	0.033	0.013	2876	2.474	0.01343 *

despite there being about 1/5 as much data, namely, 1 unit of activity compared to 5 units on average contributing to the before and after counts. The effect of doing outside the target unit is stronger when more doing occurs before the unit (0.143) than after the unit (.078). This difference is consistent with prerequisite relationships between units. In other work, we are exploring how a combination of a finer grain analysis (estimating effects of every unit on every other unit) and text processing might produce a yield discovery or validation of prerequisite links between units.

Interestingly, there is still a significant effect for doing after the target unit. This effect suggests that, in addition, to possible causal effects of within and before unit practice on improved outcomes, there may also be some general student trait (e.g., a conscientious doer/learner) that yields more practice and more learning (through some mechanism not captured in the observed data, such as greater mental effort).

For watching video, the before effect is strongest, within next, and after. Perhaps watching the videos is important to produce a level of sense making that better enhances preparation for learning from future units. Nevertheless, we once again see a much stronger effect on learning of doing than reading or watching. In fact, the ratios are even bigger with the effect of doing being 18 times more than the (non-significant) effect of reading and 7 times more than the effect of video watching.

3. TESTING THE GENERALIZATION OF THE DOER EFFECT TO OTHER COURSES

To evaluate whether the *doer effect* generalizes to other courses, we analyzed data from OLI for four courses: Computing, Biology, Statistics, and Psychology. In the case of Psychology this is the same OLI content (online readings and interactive activities) as in the MOOC, but this data comes from a different student population (those enrolled in a

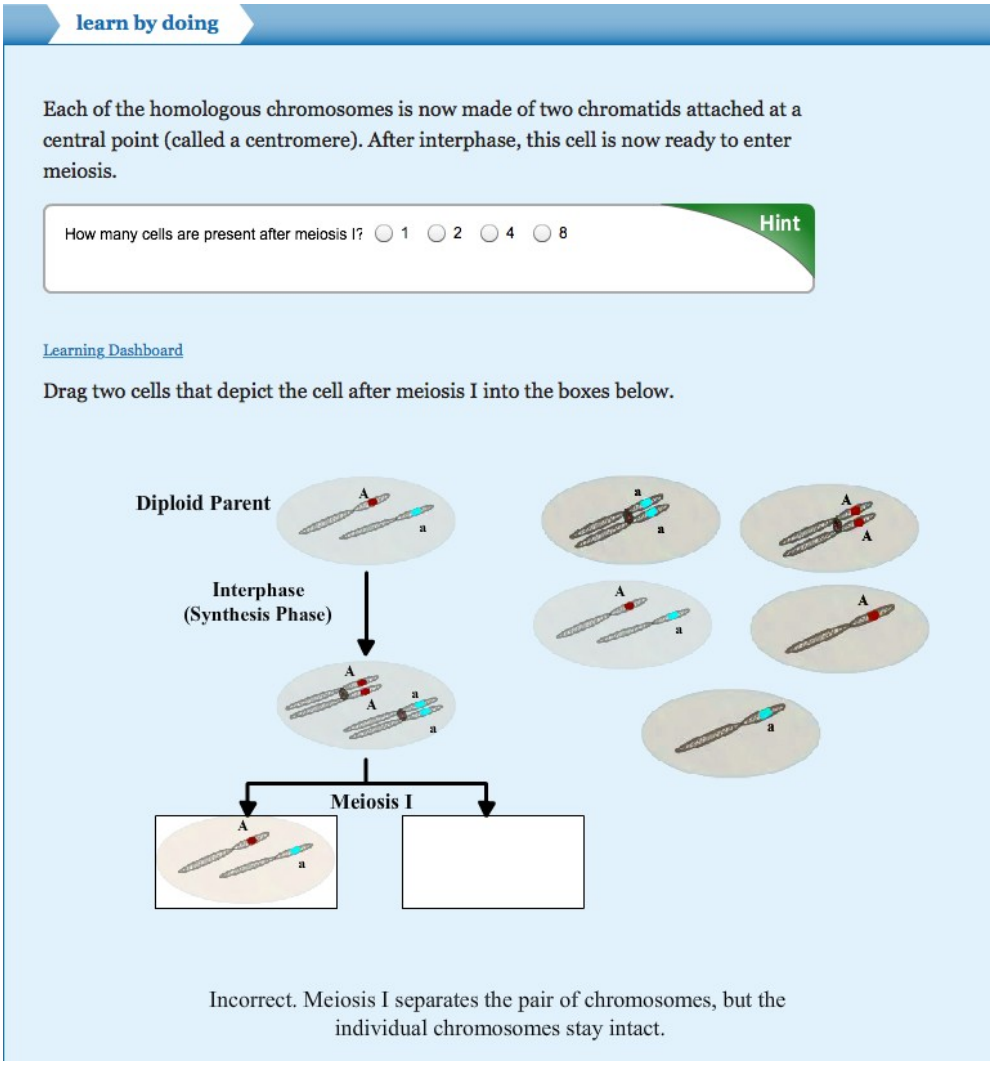


Figure 2. Biology learn-by-doing activity supporting classical genetics learning outcomes. Such activities often involve multiple related steps and integrate different forms of interaction, such as the radio box multiple choice at the top and the drag-and-drop selection at the bottom. Students get immediate feedback on their entries and can ask for hints if they are stuck.

university course rather than in a MOOC) and does not involve the video content found in the Coursera course. We evaluated how student choices to do activities and read (or at least access) pages were associated both with their total quiz scores and their course final grade. Given the similarity in surface characteristics between the online activities and the quizzes, these provide a kind of near transfer assessment of learning. The final grade is a more subjective and a more coarse measure of student learning involving more instructor judgment and making fewer distinctions between students as there are only 5 levels (A, B, C, D, or F). However, final grade has the benefit of assessing learning more broadly and serves as an intermediate (if not far) transfer assessment of learning.

3.1 Method

The University of Maryland University College (UMUC) ran a study to examine the effect of OLI resource materials on distance learning. OLI log data was collected from six courses in four disciplines (2 biology courses, 2 statistics courses, 1 computing course, 1 psychology course). Demographics (e.g., age, race, gender) were mostly evenly distributed in all classes. Inclusion criteria for our analysis consisted of (1) OLI registered students (i.e., non-OLI class sections were excluded) and (2) only students who completed the course and received a final grade (i.e., students who withdrew, failed due to non-attendance, got an incomplete, etc. were excluded).

Table 4 shows some general characteristics of the courses involved. They all have a high number of available interactive activities and readings and in all cases students tend to use a substantial number of them. Figure 2 shows an example of an interactive activity in the biology course.

Table 4. Characteristics of course use showing substantial activity use and variation.

	Students using online materials	Activities available	Activities done mean (stnd dev)	Readings available	Readings done mean (stnd dev)
Information Systems	7739	153	52 (43.7)	151	94.4 (88.8)
Biology	4564	544	200 (133.8)	881	571 (416.4)
Statistics	359	428	312 (116.0)	441	688 (420.2)
Psychology	123	687	621 (114.5)	545	510 (236.9)

3.2 Results

Figure 3 provides a scatterplot of the total pageviews (reading estimate) and total activities started (doing estimate) for the Biology course. It illustrates that while there is a positive association between reading and doing (R^2 = .342), there is also variation with some students doing more and reading less and others reading more and doing less. The scatterplots for the other courses are similar. The triangular white space on the right in the scatterplot illustrates that students must access a minimal number of pages to reach the number of activities they do. Thus, as mentioned above, to improve the estimate of reading we adjusted each student's reading score by subtracting this minimum (computed as a ratio of the activities the student did).

To pursue the question of whether the doer effect generalizes to the UMUC data, we ran regression models across the courses to assess how strongly student differences in resource use were associated with learning outcomes. These models are simpler than those in [7] to facilitate a uniform approach across datasets and because some data is not available (e.g., video watching and pre-test results) in most datasets. We performed two linear regressions for each course, one where the outcome variable was students' total quiz score and another where it was their final grade converted to numeric score (F = 1 to A = 5). We converted predictor and outcome variables to Z scores as above. The R calls used can be summarized as follows (where the brackets [ ] indicate options selected in 10 separate calls):

lm([totalQuiz.z, final_grade_in_number.z] ~
activities.z + non_activities_reading.z,
data = [Statistics, Biology, IFSM, Psych, Psych MOOC])

The results are shown in Table 5. As shown, the doer effect is consistently observed. The standardized coefficient of the effect of doing on outcomes is always significant and much higher than the standardized coefficient of the effect of reading (not always significant). The ratio of the size of the doing to reading effect goes from 2.2 to infinity (because in one case, quiz score in Statistics, the reading effect is not positive) with median of 6 -- the same ratio we found previously! In other words, the effect of doing is generally about 6 times greater than the effect of reading across four different courses and involving over 12,500 students.

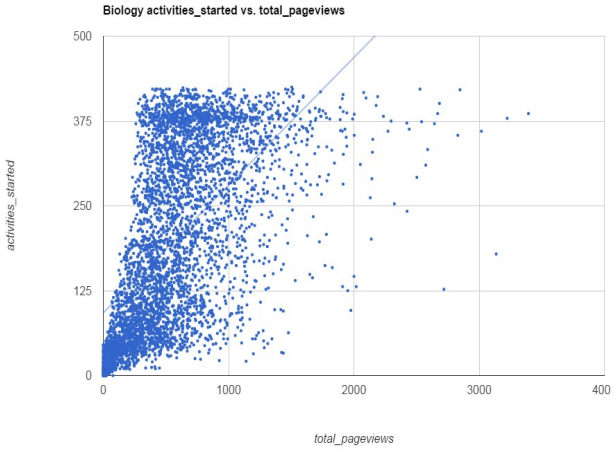


Figure 3. Scatterplot of pageviews (reading) and activities started (doing) by 4564 biology students showing that many students do more and read less (top left) and others read more and do less (lower right and middle).

Table 5. Model fit, standardized coefficients, and doer effect ratio for 5 courses and 2 outcomes.

	Quiz				Final Grade			
	Adj R^2	**Doing std coef**	**Reading std coef**	**Effect ratio**	**Adj R^2**	**Doing std coef**	**Reading std coef**	**Effect ratio**
InfoSystems	0.49	0.642	0.124	5.2	0.08	0.227	0.105	2.2
Biology	0.39	0.571	0.114	5.0	0.16	0.340	0.109	3.1
Statistics	0.24	0.519	-0.127	∞	0.11	0.327	0.020	16.4
Psychology	0.64	0.781	0.092	8.5	0.45	0.654	0.085	7.7
Psy MOOC	0.25	0.467	0.069	6.8	0.08	0.259	0.054	4.8

4. DISCUSSION AND CONCLUSIONS

Determining causal relationships is important for scientific and practical reasons because causal relationships provide a path toward explanatory theory as well as reliable and replicable practical application: If we can be certain a learning method is causally related to more optimal learning, then that method should be used to guide course design and students should be encouraged to use it. There are lots of laboratory experiments of the "testing effect" that provide high internal validity support that, in the content and contexts sampled, there is a strong causal impact of doing on longer-term learning. The content in these studies has typically been facts (even arbitrary associations involving non-words like "zep with house" [14]) and the assessments of learning have typically involved delayed retrieval of those facts. Even in the classroom studies, the orientation has been toward facts and other forms of verbal expression of concepts. Given the importance of skills and principles in many domains and given other experimental results in such domains that point to less testing and more study [4, 6, 15], it is critical that we expand efforts to test the generalizability of learning by doing. One of the strengths of these results is that the doer effect is demonstrated across four different content domains (information systems, biology, statistics, and psychology). One of these is in the humanities (psychology) but none are in the arts and it is worth investigating whether the doer effect is found in less well-defined domains, such as law or design.

Such efforts are particularly important in the context of MOOCs where so much emphasis has been placed on online lecture video. We have identified a "doer effect", an association between more doing and more learning, in data from multiple online courses. We have also shown that this effect cannot be explained solely by some global student trait, a particular third variable alternative to a causal explanation (e.g., a motivation to both do and learn). Such an explanation does not predict that, for the same student, within-unit activity will predict learning on unit content above and beyond outside-unit activity. Of course, other third variable explanations are still possible (e.g., interest in a particular unit content produces more doing and more learning) and experimentation is warranted, especially as so-called A/B testing is becoming easier to do online.

MOOC providers and online course developers should not only be pushing to be sure to have a large volume of activities, but to provide guidance and incentives to students to do them. Further, they should be exploring what are the best ratios of active doing to passive study through reading text or watching lectures? Analytics can help. We suspect that detailed online course data of the kind we analyzed can inform this question. In particular, one can investigate, for a fixed student time allocation, what ratio of doing to study is associated with the most learning.

5. ACKNOWLEDGMENTS

This work was supported by a National Science Foundation grant (ACI-1443068) toward the creation of LearnSphere.org and by funding from Google.

6. REFERENCES

[1] Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research on teaching.* Chicago: Rand McNally

[2] Dewey,J. (1916), (2007 edition). *Democracy and Education.* Teddington: Echo Library.k.

[3] Hill, P. (2013). Emerging Student Patterns in MOOCs: A (Revised) Graphical View. *e-literate*. Available online: http://mfeldstein.com/emerging-student-patterns-in-moocs-a-revised-graphical-view/.

[4] Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 75–86. doi:10.1207/s15326985ep4102_1

[5] Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8,* 30-43.

[6] Koedinger, K. R., Corbett, A. C., & Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36 (5)*, 757-798. doi:10.1111/j.1551-6709.2012.01245.x

[7] Koedinger, K. R., Kim, J., Jia, J., McLaughlin, E. A., & Bier, N. L. (2015 Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC. In *Proceedings of the Second (2015) ACM Conference on Learning at Scale*, 111-120.

[8] McDaniel, M.A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in theclassroom. *European Journal of Cognitive Psychology, 19*(4/5), 494-513. doi:10.1080/09541440701326154

[9] National Research Council. (2002). *Scientific research in education.* Committee on Scientific Principles for Education Research. Shavelson, R.J., and Towne, L., Editors. Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

[10] Pane, J.F., Griffin, B., McCaffrey, D.F. & Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis, 36* (2), 127 - 144. doi:10.3102/0162373713507480

[11] Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing Instruction and Study to Improve Student Learning (NCER 2007-2004)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

[12] Pavlik, P. I., Jr., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science, 29*, 559-586.doi:10.1207/s15516709cog0000_14

[13] Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: the effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23,* 90–108. doi:10/1006/ceps.1997.0959

[14] Roediger, H. L. & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181-210. doi:10.1111/j.1745-6916.2006.00012.x

[15] Salden, R.J.C.M., Koedinger, K.R., Renkl, A., Aleven, V., & McLaren, B.M. (2010). Accounting for beneficial effects of worked examples in tutored problem solving. *Educational Psychology Review*. doi: 10.1007/s10648-010-9143-6

[16] Singer, S.R. & Bonvillian, W.B. (2013). Two Revolutions in Learning. *Science 22*, Vol. 339 no. 6126, p.1359. doi:10.1126/science.1237223

[17] Lovett, M., Meyer, O., & Thille, C. (2008). The Open Learning Initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education, 2008* (1), 1-16. http://doi.org/10.5334/2008-14

[18] Trochim, W. M. (2009). Evaluation Policy and Evaluation Practice. *New Directions for Evaluation, 123*, 13-32.doi:10.1002/ev.303

[19] Wang, X., Wen, M., Rosé, C. P. (2016). Towards triggering higher-order thinking behaviors in MOOCs, in *Proceedings of Learning, Analytics, and Knowledge* '16.

[20] Wieman, C. E. (2014). Large-scale comparison of science teaching sends clear message. *Proceedings of the National Academy of Science, 111*(23), 8319 – 8320. doi:10.1073/pnas.1407304111

[21] Zhu, X., & Simon, H.A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction*, 4, 137- 166. doi:10.1207/s1532690xci0403_1