The Doer Effect: Replication and Comparison of Correlational and Causal Analyses of Learning

Rachel Van Campenhout & Benny G. Johnson
Research and Development
VitalSource Technologies
Pittsburgh, USA
Email: rachel.vancampenhout@vitalsource.com

Jenna A. Olsen
Learning Analytics
Western Governors University
Salt Lake City, USA
Email: jennaanneolsen@gmail.com

Abstract - **The doer effect is a learning science principle that proves students who engage with formative practice at the point of learning have higher learning gains than those who only read expository text or watch video. This principle has been demonstrated through both correlational and causal analysis. It is imperative that learning science approaches capable of increasing student learning gains be rigorously tested and replicated to confirm their validity before wide-scale use. Previously we replicated causal doer effect results using student data from courseware used at a major online university. In this paper, we will replicate both the correlational doer effect analysis as well as the causal analysis using both unit tests from the courseware and the course final exam. These multiple analyses of the doer effect on the same course data provide a unique comparison of this method and the impact of the doer effect on near and intermediate learning assessments. Findings of the correlational doer effect analyses confirmed doing was more significant to outcomes than reading, and further analysis determined these results could not be attributed to student characteristics. Results of the causal analysis verified doing was causal to learning on both the unit tests and final exam. The implications of these doer effect replication results and future research will be discussed.**

Keywords - doer effect; learn by doing; causal discovery; replication; external validity; learning outcomes; course effectiveness; courseware.

I. INTRODUCTION

Students deserve digital learning resources that actually help them learn. And yet, verifying which methods are effective for learning is a challenging task. A benefit of courseware as a comprehensive learning environment is the wealth of data available for analysis. As students move through the courseware, their page visits, engagement and accuracy on formative practice, summative assessment scores and more can be collected to paint a picture of what students are doing both in real time and for post hoc analysis. The large-scale data from courseware run in natural settings can be used as a basis for investigating the effectiveness of learning methods. The courseware data can provide many insights, if the right questions are asked. One such question is: Are we able to identify if courseware's formative practice questions cause increased learning?

The doer effect is the learning science principle that the amount of interactive practice a student does (such as answering practice questions) is much more predictive of learning than the amount of reading or video watching the student does [11]. Studies have shown correlational support for this principle [10]. The total amount of reading and total amount of doing are used in a linear regression to identify the doer effect coefficient as a means of quantifying the doer effect. Koedinger et al. [11] found that doing had a median of six times the relationship to learning than reading.

However, in order to recommend this approach with high confidence in its effectiveness, it is necessary to know that there is a causal relationship between doing practice and better learning. This requires ruling out the possibility of a third variable being a common cause of both, since in that case the relationship between doing and learning would merely be correlational. For example, a frequently cited external variable that could account for the doer effect is student motivation. A highly motivated "go-getter" student may do more practice and also obtain better learning outcomes, but this would not necessarily mean better outcomes were *caused* by doing the practice.

Koedinger et al. [10] used data collected from students engaged with a MOOC course paired with courseware developed by Carnegie Mellon's Open Learning Initiative (OLI) to investigate the doer effect. In their initial research, they found the learning effect of doing the formative practice was about six times larger than that of reading. Follow-up analysis [11] [12] sought to determine whether this effect was causal. A statistical design involving within- and outside-unit reading, watching and doing (described in more detail below), was able to demonstrate causal impact of doing on learning and rule out the possibility that this effect was entirely the result of a factor such as individual student motivation. There is no better explanation of the importance of causal relationships than was stated in [11]: "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."

Replication research is critical in the learning sciences to provide additional evidence to support—or refute—claims made about effective learning practices. A large fraction of published research in the social sciences has not been replicated, and studies that cannot be reproduced are cited

more frequently than those that can [16]. Methods for increasing learning should be broadly shared to benefit as many students as possible, and those methods should be grounded in substantial evidence of their validity.

Reproducible research is not only necessary for the research community, but for practical application in educational technology. The courseware analyzed in this study was developed using the methods and approach of learning engineering—a practice that supports learners and their development through the application of the learning sciences to human-centered engineering design methods and data-driven decision making [7]. Proposed by Herbert Simon [17] and fostered at Carnegie Mellon University [6], learning engineering developed as a role to further the application of learning science for students and instructors. Learning engineering was applied at Acrobatiq after its emergence from OLI to apply learning science and a student-centered approach to developing courseware [19]. The Learning Engineering Process (LEP) outlines an iterative cycle that includes the identification of the context and problem, design and instrumentation, implementation, and data analysis and results [8]—a development process appropriate for many contexts. While the application of learning science research was a critical component of the LEP for the development of the courseware, equally vital is the analysis of data and sharing results. To fully engage the LEP is to iteratively improve through the insights data can reveal, and to share these findings with the broader research community. A goal of this paper is to further the LEP by collaborating with an institutional partner to replicate learning science research foundational to the courseware through the analysis of data gathered from students in a natural learning context. By replicating and sharing the data analysis and findings as part of the LEP, the researchers and developers maintain transparency and accountability to the learner [19].

Furthermore, replicating findings that are based on large-scale data mining provides valuable verification of the results, as the volume and type of data analyzed can be difficult to obtain. Through the courseware described in this paper and institutional collaboration, we have the data required to evaluate the relationship between doing practice and learning outcomes. Replicating this causal doer effect study adds to the body of evidence that this learning by doing methodology—and the doer effect it produces—are effective in a variety of learning situations, and supports a practical recommendation that students can increase their learning outcomes by increasing the amount of formative practice they do.

For this study, the data set came from students enrolled in a Macroeconomics course, C719, at Western Governors University. There are many benefits of analyzing student data from courseware used in a real university setting. Students engaged with the course without any external influences that might alter their natural behavior. This allows us to study their engagement and learning outcomes in as authentic a way as possible; students worked through this course as they would any other in their program, which contributes to the generalizable nature of the study. Benefits of utilizing real course data include lower costs and fewer ethical concerns as compared to controlled experiments. A controlled experiment in a laboratory setting would allow researchers to, for example, deliver the treatment (doing practice interleaved with content) to one randomly selected set of students while delivering static content to a control group. Performance on a standard assessment would provide a measure of the effect of the treatment. This controlled experimental method would have a high internal validity, but would also have a high cost, ethical concerns, and low external validity. Instead, due to the availability of detailed data generated by courseware as students progress through their course, post hoc studies of natural learning contexts can be done with minimal cost and without ethical concerns that can come with randomized experiments, such as withholding potentially beneficial treatment from some learners.

The value of this replication study is that it extends the external validity of the doer effect findings. The Macroeconomics courseware used was designed on the Acrobatiq platform based on the principles established at OLI. This courseware utilizes the same key features of interleaved practice, immediate targeted feedback, etc., as the OLI courses previously analyzed (Introduction to Psychology, Introduction to Biology, Concepts in Computing, Statistical Reasoning) [11]. These similarities are important for confirmatory results, as it is important to have as many common variables as possible for the replication of the statistical model [12]. Investigating courseware in an entirely different subject domain built independently—yet using the same learning science principles—strengthens the external validity of a causal relationship.

This study extends our previous doer effect replication research [1] by replicating Koedinger et al.'s [11] correlational and causal doer effect analysis, using both the unit test summative assessments from the courseware as well as the WGU final exam. This analysis provides a direct comparison of both the correlational and causal analysis on the same course, providing insight into the comparison of outcomes between assessment types. Additionally, demographic information collected by WGU will be used to extend the correlational model—as done in Koedinger et al. [10]—to investigate how additional variables impact the doer effect coefficient.

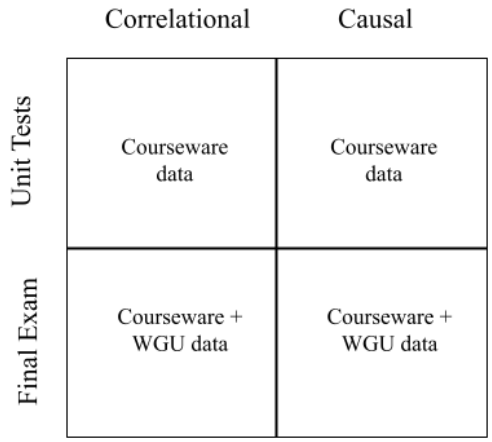

Figure 1. The doer effect analyses in this paper.

Given the intention of this study to replicate doer effect findings—both correlational and causal—our research questions are:

1. Can the correlational doer effect be replicated using both courseware unit tests and final exam scores?
2. Can the doer effect be accounted for by student characteristics?
3. Can the causal doer effect be replicated using both courseware unit tests and final exam scores?

To answer these questions, we will outline the required parallel features for this replication study in Section II—from the learning by doing courseware environment, to the description of the regression model and its inputs. Section III will provide the methods, results, and discussion on the correlational doer effect analyses. Section IV will outline the methods, results, and discussion for the causal doer effect analyses. Section V concludes the paper with remarks on the importance of these replication findings for the learning science methods used herein, and the implications of these findings for future research.

II. STUDY 1: CORRELATIONAL DOER EFFECT REPLICATION

This section will provide the methods, results, and discussion for the correlational doer effect models, using both the unit tests and final exam as the outcome. This section also includes additional analysis of the correlational doer effect ratio when controlling for student characteristics.

A. Methods

In order for this replication research to be parallel with the original study [11], the learning resource needed to be similar in the learning by doing approach. The term "learning by doing" has been broadly used to describe various kinds of learning engagement (and not all use or encourage the use of scaffolding or feedback [9]), so it is important to clarify how learning by doing is applied in this courseware. Learning by doing is a method of actively engaging the learner in the learning process by providing formative practice at frequent intervals. It has been shown that formative practice increases learning gains for students of all ages and in diverse subjects, and while this method benefits all students, it can benefit low-performing students most of all [4]. The formative practice questions integrated with the content essentially act as no-stakes practice testing, which increases learning gains and retention [5]. In Acrobatiq courseware, students can answer practice questions as many times as they like, and typically students continue to answer until they get the correct answer [20]. Feedback that explains why that choice is correct or incorrect is provided for each answer option to give additional guidance and another opportunity for learning (Figure 2). Immediate, targeted feedback was shown to reduce the time it took students to reach a desired outcome [2] [13], and feedback in practice testing outperforms no-feedback testing [5] [15]. Formative practice with targeted feedback provides scaffolding and examples that support cognitive structures for effective learning [9] [15] [18].

The courseware contains many features similar to those used in the courses for the original study [11]. Modules are made up of lesson pages, and each lesson contains readings, images, and formative practice questions all tied to a central learning objective. Learning objectives are student-centered and measurable, and the practice questions are tagged with the learning objective to feed data to the platform's learning analytics engine [21], as well as to inform post hoc analysis. The formative practice questions are interleaved with small chunks of content to provide practice to students at the point of learning that content. Question types vary but entail both recognition and recall and most frequently include multiple choice, pulldown, text or numeric input, drag and drop, and true/false. Questions were created to target the foundational Bloom's Taxonomy category, *remembering*, of which recognition and recall are both cognitive processes [3] [20].

In addition to formative practice questions integrated within the content, there are adaptive activities and summative assessments in the courseware [21]. The adaptive activities are placed at the end of the module and cover all learning objectives included in that module. The questions in the adaptive activities are personalized to the needs of each student at the time they enter the activity. Student performance on the formative practice generates a learning estimate for each learning objective, which is used to determine the scaffolded questions students receive in the activity. The adaptive activities are also formative in nature, so students received immediate, targeted feedback and could make multiple attempts. The questions on these activities are not scored. At the end of the module, students have a quiz—comprehensive to all learning objectives included in that module—that produces a grade. At the end of the unit is a unit test that includes questions on all learning objectives from all modules and also produces a grade. While these unit test grades were not included as part of the course grade (see below), students see their scores in the courseware. These unit tests are used as the summative assessments for the courseware-based doer effect analyses.

By partnering with Western Governors University, the doer effect analysis can also be done using data from the final exam. Students enrolled in the course were able to review the course content (the courseware) and work with faculty at their own pace in preparation for a final exam that comprised 100% of the course grade. Students had a six-month window to complete the course by passing the final exam, which they could retake as needed during that time frame. This learning science-based courseware was developed to fit WGU's curriculum needs.

Passing the WGU course depended solely on passing the final exam. The courseware content and final exam content were written by independent development teams; however, the course learning objectives were provided to the WGU final exam development team for alignment purposes. For the final exam-based doer effect analyses, the student's score on the first attempt at the final exam was used as the learning outcome.

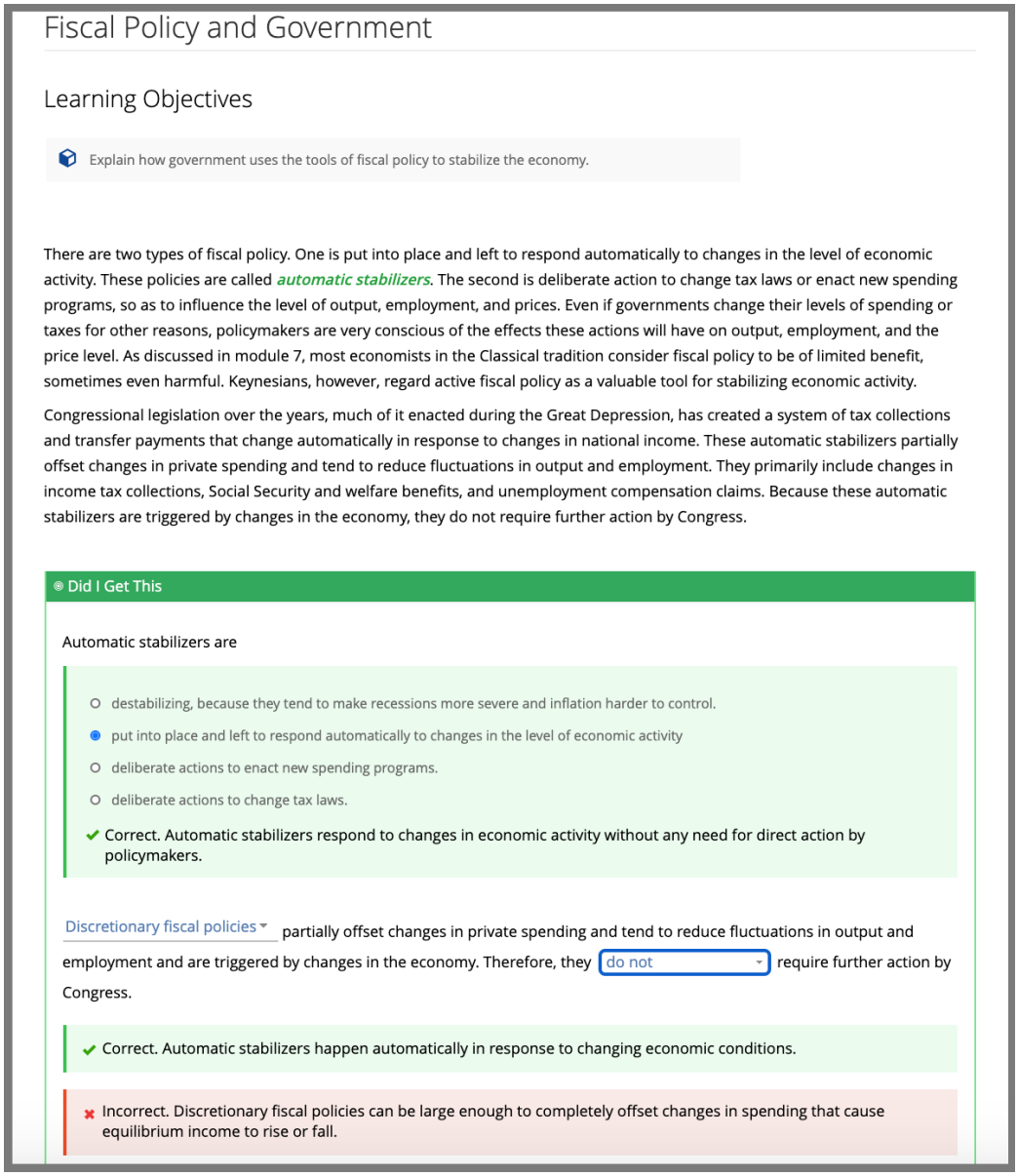

Figure 2. A lesson page with formative questions from Macroeconomics.

The model used by Koedinger et al. [11] to determine the doer effect coefficient is an ordinary linear regression that expresses the assessment outcome as a function of total reading and total doing. The ratio of doing and reading coefficients from the model determines the overall doer effect coefficient. Following Koedinger et al. [11], the reading variables were defined as all visits to lesson pages where the student did not engage in any practice available on that page. The doing variables were defined as the number of formative practice opportunities a student attempted, including in the adaptively generated practice activities described earlier. The courseware's module quizzes and unit tests were not included as practice because of their presentation as scored summative assessments, even though in this case they made no contribution to the student's grade in the course; inclusion of these as practice did not materially affect the results of the analysis. Unlike in some of the previous studies in which video lectures were used [11], video watching was not investigated here, as video was not a critical component of the courseware.

The following analyses use different assessments as the outcome (unit tests or final exam) and therefore the selection criteria for each analysis will have different numbers of students, as depicted in Figure 3. The initial data set included historical data from 3,513 students who enrolled in the Macroeconomics course from March 2017 to April 2019 (WGU courses have rolling enrollments). There are selection criteria relevant to both unit test and final exam analyses. Only students who completed the course (defined here as taking the final exam) were included. As the study we intend to replicate included only students who made some use of the course materials, we likewise excluded students who did not use the

courseware at all. WGU allowed students to take the course's final exam more than once (if necessary) to pass. Only the first attempt at the final exam was included in the analysis, and student engagement with the courseware was filtered to include only that which occurred before the first attempt at the final exam. This resulted in 3,120 students in the final exam data set. For this group of students, there were 224,072 total reading page visits and 1,143,601 total first attempts on 1,162 available formative practice questions. However, for the courseware unit test analyses, to be consistent with the selection criteria of the original study, only students who completed nearly all of the assessments (in this case at least 5 of 6 unit tests) were included. This resulted in a smaller subset of data that included 493 students.

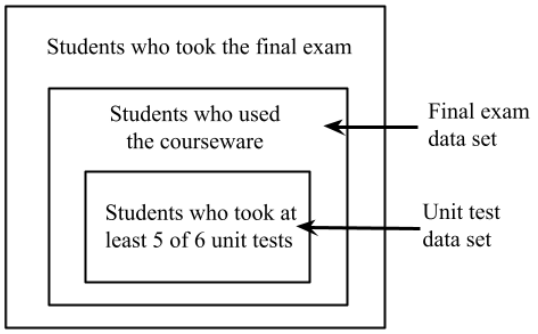

Figure 3. The subsets of student data used for the analysis.

B. Results

The first analysis uses data solely from the courseware platform. Using the reading and doing data from the courseware, we can use a linear regression to calculate the doer effect coefficient on the courseware unit tests. The model to compute the doer effect coefficient is replicated from Koedinger et al. [11], so we follow the same procedure for this analysis. The unit test data set of 493 students described above was used. Following [11], since some students did not take all unit tests, the total of the unit test scores was used as the outcome. Only reading and doing relevant to the assessments taken was included. Reading, doing, and score values were converted to *z*-scores before regression to better enable comparison of the reading and doing effects, since reading and doing are measured in different units (pages visited vs. questions answered). The R formula for the regression model is:

```
lm(z_total_unit_test_score ~ z_total_reading
                           + z_total_doing,
                             data=df)
```

The results in Table I show that the standardized reading coefficient is not significant ($p = 0.838$) while the standardized doing coefficient is highly significant ($p < 0.001$). The doer effect coefficient is the ratio of doing to reading. Previous work by Koedinger et al. [10] [11] found the effect of doing on outcomes was about six times greater than reading. In cases

TABLE I. UNIT TEST CORRELATIONAL DOER EFFECT REGRESSION ANALYSIS.

	Estimate	**Standard Error**	**t-Value**	**Pr(>|t|)**
(intercept)	0.0000	0.0403	0.000	1.000
Total Reading	0.0088	0.0429	0.205	0.838
Total Doing	0.4472	0.0429	10.420	<2e-16 ***

where Koedinger et al. [11] could not compute a size for the doer effect because reading was not significant or negatively significant, they reported such cases as an effect ratio of ∞. In this case, because reading is not significant, the confidence interval of the reading coefficient includes zero. Therefore, the doing coefficient is effectively divided by zero, giving a ratio of ∞.

As was done with the unit tests, we can use a linear regression to determine the doer effect coefficient with the final exam score as the outcome. There are 3,120 students in this data set, which is considerably more than in the previous analysis of the courseware unit tests due to the change in selection criteria. Because the unit tests in the courseware were not required to pass the course, fewer students completed them, whereas all students completing the course had to take the final exam.

```
lm(z_final_exam_score ~ z_total_reading
                      + z_total_doing,
                        data=df)
```

TABLE II. FINAL EXAM CORRELATIONAL DOER EFFECT REGRESSION ANALYSIS.

	Estimate	**Standard Error**	**t-Value**	**Pr(>|t|)**
(intercept)	0.0000	0.0170	0.000	1.000
Total Reading	-0.1069	0.0209	-5.105	3.51e-07 ***
Total Doing	0.3655	0.0209	17.450	< 2e-16 ***

The results of this model using the final exam in Table II show that total reading is negative and significant, while total doing is positive and significant. Because total reading was negative, following Koedinger et al. [11] the doer effect in this case is also reported as ∞.

In their initial research on the doer effect, Koedinger et al. [10] wanted to verify that certain student characteristics weren't accounting for the doer effect results. They created a linear regression model that accounted for: pretest score, Quiz 1 score, occupation, age, education and gender. The only significant coefficients in the model were the Quiz 1 score and education, with the OLI courseware usage still significant to final exam scores. No other student characteristics were significant.

Through the student data WGU collects, we were able to do a similar analysis to determine if any of the available student characteristics were significant to student outcomes, especially to the extent that they accounted for the doer effect findings. The student characteristics recorded by WGU were gender, underrepresented status, first-generation status, Pell eligible status, and age. These covariates were added to the linear regression model along with total reading and total doing. The R formula used for the unit test as the assessment was:

```
lm(z_total_unit_test_score ~ z_total_reading
                           + z_total_doing
                           + male + underrep
                           + first_gen
                           + pell_eligible
                           + c_age,
                             data=df)
```

The results of this linear regression are in Table III. The same model was also fit using the WGU final exam data set. The results of that model are in Table IV.

TABLE III. UNIT TEST CORRELATIONAL DOER EFFECT REGRESSION ANALYSIS INCLUDING STUDENT CHARACTERISTICS.

	Estimate	Standard Error	t-Value	Pr(>|t|)
(intercept)	0.1017	0.0673	1.513	0.131
Total Reading	0.0511	0.0413	1.237	0.217
Total Doing	0.4355	0.0407	10.710	< 2e-16 ***
Male	0.4525	0.0842	5.373	1.2e-07 ***
Under-represented	-0.4007	0.1030	-3.888	0.000115 ***
First Generation	-0.2246	0.0781	-2.875	0.00422 **
Pell Eligible	-0.1927	0.0816	-2.360	0.0187 *
Age	-0.0029	0.0042	-0.683	0.495

TABLE IV. FINAL EXAM CORRELATIONAL DOER EFFECT REGRESSION ANALYSIS INCLUDING STUDENT CHARACTERISTICS.

	Estimate	Standard Error	t-Value	Pr(>|t|)
(intercept)	0.0831	0.0296	2.810	0.00498 **
Total Reading	-0.0789	0.0205	-3.843	0.000124 ***
Total Doing	0.3445	0.0203	16.969	< 2e-16 ***
Male	0.3037	0.0349	8.697	< 2e-16 ***
Underrepresented	-0.4214	0.0406	-10.377	< 2e-16 ***
First Generation	-0.1569	0.0338	-4.642	3.59e-06 ***
Pell Eligible	-0.0986	0.0351	-2.857	0.00430 **
Age	0.0051	0.0018	2.764	0.00575 **

C. Discussion

The doer effect coefficient was ∞ for both the unit test and final exam regression model. The total doing estimate was positive and highly significant for both models. The total doing estimate was slightly smaller for the final exam compared to the unit test. This may be related to the proximity of the assessment being used as the outcome. Doing practice may have a slightly stronger effect on the outcomes of the unit test immediately following each unit as opposed to the final exam at the end of the coursework.

For the unit tests model, the reading coefficient was not significantly different from zero, and therefore the ratio was ∞. For the final exam model, the reading coefficient was negative and significant, producing the same result. Whereas the total doing estimate became slightly smaller but still significant from the unit test to the final exam model, the total reading coefficient went from not significant to negatively significant from unit test to the final exam model. This trend is interesting when comparing the results from a near proximal assessment to a distant one, and investigating this trend on additional courses would be a valuable future study.

The correlational doer effect analyses show that doing practice has a much larger effect size on learning outcomes than reading. However, as these are the only two variables evaluated as of yet, there may be a question of whether these results could be due to another variable, such as prior knowledge or demographics. The student characteristics regression models for both unit tests and the final exam show that total doing was still significant even when controlling for student characteristics. Total reading was not significant for the unit tests, and was negative and significant for the final exam score. The total doing and total reading results for the student characteristics models mirror those from the unit test and final exam correlational models. The doer effect coefficient of doing over reading can also be calculated using these student characteristics linear regressions. Just as in the previous unit test correlational model, the reading coefficient was not significant and therefore the coefficient was ∞. The results of the final exam score model that controls for student characteristics also had a negatively significant reading estimate, so the doer effect ratio is ∞ in this case as well.

The unit test and final exam student characteristics models shared similarities on which covariates were significant. In both models, being male was positively significant. Underrepresented status, Pell eligible status, and first-generation status were all negatively significant in both models. Age was not significant in the unit test model but was on the final exam model. While interpreting the trends in demographic characteristics is outside the scope of this paper, what is key from this analysis is that the doer effect is still present even when controlling for student characteristics. A recent cluster analysis study on data from Western Governors University sought to understand which student characteristics provided the most value for predicting student success [14]. Results found that student activity attributes were more valuable as a stand-alone category, exceeding the value of both student readiness and demographics. Combined with the verification that the doer effect ratio remains unchanged when controlling for student characteristics, this suggests that future research should focus on how to increase engagement with formative practice for all students to maximize the benefits of the doer effect for all.

III. Causal Doer Effect Replication

This section will provide details on the methods, results, and discussion of causal doer effect analyses using both unit tests and the final exam.

A. Methods

The correlational doer effect results obtained through the traditionally used ordinary linear regression approach are useful and informative; however, the old maxim "correlation is not causation" still applies. We earlier discussed the importance of whether the relationship between doing and outcomes is in fact causal, not merely correlational. Is it possible to go beyond a correlational model and answer this key question? For this purpose, Koedinger et al. [11] developed a regression model that analyzed the relationship of student doing, reading, and video watching in each unit of course content to scores on that unit's summative assessment. The key innovation in their model was to control for the total amounts of doing, reading, and watching in *other* units of the course. Student doing outside the unit can act as a proxy for a third variable like motivation that can lead to correlation between level of effort and outcomes. In this way, if the doer effect is causal, then the amount of doing within a unit should be predictive of the student's score on that unit's assessment, even when accounting for doing outside that unit. If there is not a causal relationship between doing and outcomes, we would not expect to see a statistically significant within-unit effect beyond the outside-unit effect.

Replicating this causal model using the unit tests from the courseware is a fairly straightforward task. As was done for the correlational models, the reading variable was identified as visits to lesson pages where students did not do practice (when available) and the doing variable was defined as attempts on formative practice. The courseware unit becomes the container for determining within-reading and within-doing, while all pages not in the unit become outside-reading and outside-doing. The unit test is the summative assessment for the unit and serves as the outcome variable for the model. Just like for the correlational model, using the unit tests as the outcome had the additional selection criteria of completing 5 of 6 unit tests, creating a subset of 493 students.

Unlike in the original study, where a summative assessment immediately followed each unit of course content, the final exam was obviously taken after all relevant student usage of the courseware. Furthermore, the units in the Acrobatiq courseware did not have a direct correspondence with the categorization of questions on the final exam. As previously discussed, all courseware resources, e.g., lesson readings and formative practice questions, were mapped to the course learning objectives. These learning objectives in turn mapped to six course competencies developed by WGU, to which final exam questions were also coded. The six competencies for the course are:

1. 3003.2.1 – The Economic Way of Thinking - The graduate analyzes economic behavior by applying fundamental economic principles, including scarcity, opportunity cost, and supply and demand analysis.
2. 3003.2.2 – Macroeconomic Measurements and Theories - The graduate analyzes unemployment, inflation, economic growth, business cycles, and related economic theories.
3. 3003.2.3 – Federal Budget and Fiscal Policy - The graduate explains fiscal policy and its effects on the federal budget, national debt, and economy.
4. 3003.2.4 – Money, Financial Markets, and Monetary Policy - The graduate analyzes the monetary system, including the influence of monetary policy on the economy.
5. 3003.2.5 – Economic Growth and Development - The graduate explains how macroeconomic policies affect economic growth and development.
6. 3003.2.6 – International Trade - The graduate explains how trade policies influence international markets.

In order to apply the Koedinger et al. regression model [11] using the WGU final exam, these course competencies were used as the analysis units, as this provided a way to group both the courseware content and the final exam questions into a common set of logical units. Thus, when referring to a *unit* of course content for the final exam model, we specifically mean all content corresponding to one of these six competencies, with the unit summative assessment consisting of all corresponding final exam questions that assess that competency.

The competencies were used to compile the unit-based reading and doing data required for the model from the clickstream usage events logged by the courseware. Within-unit resource use (reading or doing) was defined as all use associated with a unit's content, and outside-unit resource use was defined as all resource use not designated as within-unit. In total, 47 finer-grained courseware learning objectives were mapped to the six course competencies. The learning objectives were not uniformly distributed across competencies, as the number varied according to the amount of content coverage. The mapping of the courseware's formative practice to the learning objectives was used to aggregate practice by competency.

B. Results

Using the unit test data set, we can replicate the causal doer effect model from [11] using the courseware summative assessments as the outcome. The same procedures in preparation for applying that regression model were followed here as well, such as confirming that there is sufficient variation in individual student reading and doing across course units for the analysis [11]. The score distribution for each of the unit tests in the courseware is shown in Figure 4. The boxplots were generated using R's boxplot function, which uses the interquartile range rule to determine outliers, shown as circles.

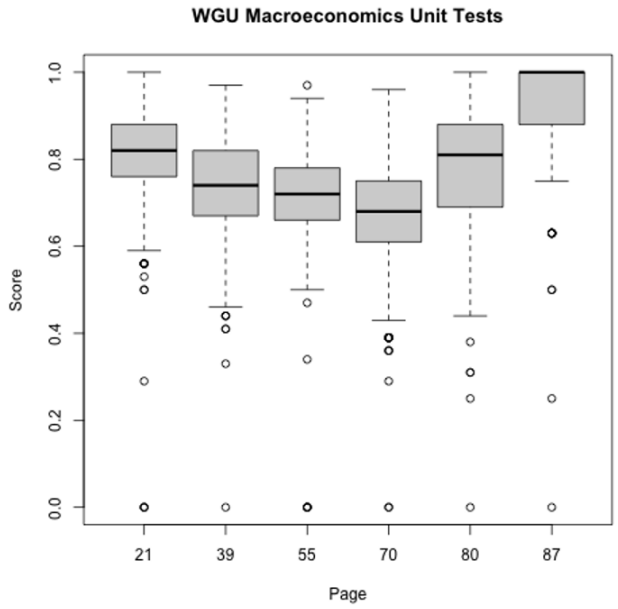

Figure 4. Score distribution for the six Macroeconomics unit tests.

Unlike in the correlational studies, here we have an observation (row) for each individual unit test each student took, giving 2,807 total observations of the 493 students in this data set. The multiple observations per student are not independent and therefore an ordinary linear regression model—which assumes independence—cannot be used. The lack of independence can be handled by using a mixed effects linear regression model. Following Koedinger et al. [11], we use a mixed effects model to investigate the within-unit and outside-unit reading and doing relationships with learning outcomes. The R formula used to fit the model is below. This shows that a linear mixed effects regression model was fit using the lmer function. The regression formula shows unit test score modeled as a function of within- and outside-unit reading and doing, with a random intercept per student and unit test to address the lack of independence of the observations noted above.

```
lmer(z_unit_test_score ~ z_within_reading
                       + z_outside_reading
                       + z_within_doing
                       + z_outside_doing
                       + (1|student)
                       + (1|unit_test),
                         data=df)
```

The reading and doing coefficients were tested for statistical significance using a likelihood ratio test, in which the likelihood of the full model is compared to a model with one of the variables of interest omitted. The R code below illustrates this test for the within-reading coefficient.

The results of the regression model in Table V show that both within- and outside-unit doing are positive and significant at $p < 0.001$. Within-unit reading is positive and significant at $p < 0.05$; however, outside-unit reading is negative at $p < 0.05$.

```
lme.model <- lmer(z_unit_test_score ~ z_within_reading + z_outside_reading + z_within_doing
                                    + z_outside_doing + (1|student) + (1|unit_test),
                                      data=df, REML=FALSE)
lme.null  <- lmer(z_unit_test_score ~ z_outside_reading + z_within_doing + z_outside_doing
                                    + (1|student) + (1|unit_test),
                                      data=df, REML=FALSE)
anova(lme.null, lme.model)
```

TABLE V. UNIT TEST CAUSAL DOER EFFECT REGRESSION ANALYSIS.

	Location	Estimate	Std. Error	t-Value	Pr(>|t|)
	(intercept)	0.0106	0.2778	0.038	0.967
Reading	within-unit	0.0483	0.0198	2.441	0.0146 *
	outside-unit	-0.0559	0.0268	-2.085	0.0370 *
Doing	within-unit	0.1629	0.0276	5.902	5.17e-09 ***
	outside-unit	0.1272	0.0271	4.697	2.70e-06 ***

Next, we can repeat this same causal doer effect analysis using the final exam as the outcome. Prior to creating the linear regression model, we examined the score distribution for each of the six competencies on the final exam, shown in Figure 5. It is seen that the competencies have differing student score distributions.

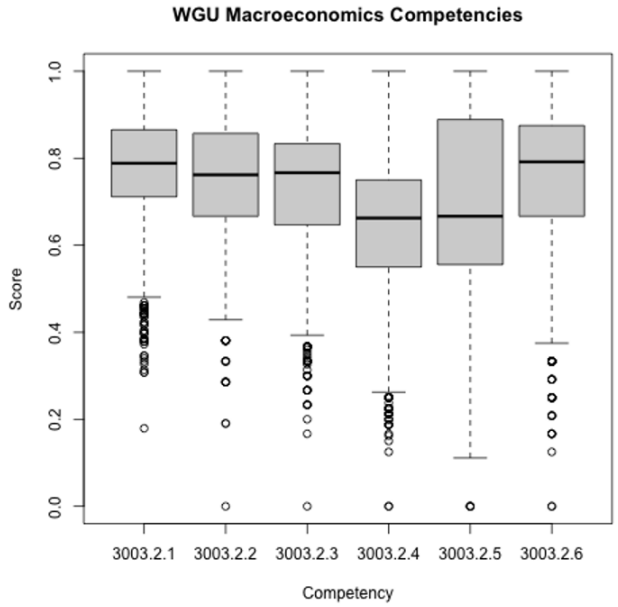

Figure 5. Score distributions for the six competencies of the Macroeconomics final exam.

For each of the 3,120 students in the data set, there is an observation for each of the six competencies, bringing the total number of observations to 18,720. As with the unit test mixed effects regression model, the model was fit with the `lmer` function, and shows competency score modeled as a function of within- and outside-unit reading and doing.

The R code for this model is:

```
lmer(z_WGU_COMPETENCY_SCORE ~ z_within_reading
                            + z_outside_reading
                            + z_within_doing
                            + z_outside_doing
                            + (1|student)
                            + (1|competency),
                              data=df)
```

The reading and doing coefficients were tested for statistical significance using a likelihood ratio test as done for the unit test model.

The results of the regression analysis are presented in Table VI. There are significant effects for within-unit doing, outside-unit doing, and outside-unit reading, while within-unit reading is not significant. The within-unit and outside-unit doing coefficients are larger in magnitude than both the reading coefficients, and doing also had much larger *t*-values than reading. The reading coefficients are also negative, which we will discuss further below.

C. Discussion

The course analyzed in Koedinger et al. [11] had eleven total content unit/assessment pairs. Within-unit doing and watching were significant, as well as outside-unit doing. Reading and outside-unit watching were not significant. Outside-unit doing significance indicates that there is a variable that influences how students who generally do a lot of practice also score higher on assessments. However, the larger and more significant predictor was within-unit doing, meaning that even when controlling for outside-unit doing, within-unit doing had a statistically significant relationship with learning outcomes, indicating a causal doer effect.

For both the unit test and final exam model, both within-unit doing and outside-unit doing were strongly, positively significant. We initially discussed how significant within-unit doing would be indicative of a causal relationship between doing practice and better learning outcomes. But since outside-unit doing is also significant, does that mean that a causal doer effect is *not* supported? No. We would likely

TABLE VI. FINAL EXAM CAUSAL DOER EFFECT REGRESSION ANALYSIS.

	Location	Estimate	Std. Error	t-Value	Pr(>|t|)
	(intercept)	0.0000	0.1256	0.000	1.000
Reading	within-unit	-0.0125	0.0091	-1.367	0.173
	outside-unit	-0.0604	0.0130	-4.645	3.43e-06 ***
Doing	within-unit	0.1146	0.0099	11.613	< 2.2e-16 ***
	outside-unit	0.1556	0.0132	11.773	< 2.2e-16 ***

expect outside-unit doing to almost always be significant (regardless of whether the doer effect is causal), as it is well known that students who do more practice tend to get better outcomes. Significance of outside-unit doing simply reflects that, for example, students who are go-getters typically do well. What matters is that within-unit doing is *additionally* significant, which means the relationship of within-unit doing to its own unit's assessment score cannot be accounted for by the amount of outside-unit doing, indicating that relationship is causal in nature. Otherwise, we would expect outside-unit doing to be significant but not within-unit doing. But this is not the case: within-unit doing matters to learning outcomes in a way that cannot entirely be explained by a third variable—such as motivation—that leads to both greater doing and better learning. The most important finding is therefore that within-unit doing is a highly significant predictor of learning even after controlling for outside-unit doing, and this is consistent with a causal doer effect. That this finding is consistent between the unit test model and final exam model provides additional confirmation that the doer effect is present on both near transfer assessments (the unit test) as well as medium to far transfer assessments (the final exam).

Within-unit reading in the unit test model had a smaller estimate than doing but was still positively significant, indicating that reading the unit content was beneficial for the unit test. Because within-unit reading was positive and significant, we can use within-unit doing to calculate the doer effect ratio for the unit tests: 3.4. While this is less than the value of 6 typically quoted as the representative doer effect size, Koedinger et al. [11] found doer effect ratios ranging from 2.2 to ∞. However, when we look at the final exam model, within-unit reading is no longer significant. This could be indicating that the within-unit reading is not beneficial for the far transfer of the final exam assessment when compared to the near transfer of the unit test. Because it is possible within-reading is not statistically different than zero, the doer effect ratio for the final exam is, once again, reported as ∞.

An interesting note is that the outside-unit reading coefficient was significant but negative on both the unit test and final exam models, showing an overall negative relationship between the amount of outside-unit reading and assessment performance. One possible explanation for this negative result is suggested from prior anecdotal observations of engagement behaviors of students with poor learning outcomes. Many of these students tended to read the same section(s) of text repeatedly, indicating they were struggling. This pattern of rereading without obtaining a good outcome may have contributed to this negative relationship. These struggling students also often did not meaningfully engage in practice, which is regrettable since the body of doer effect research would recommend that investing that study time in practice instead of rereading would have been more beneficial. Note particularly that within-unit reading was not significant, meaning no special relationship to outcomes beyond outside-unit reading was discernible. This negative relationship between reading behavior and outcomes should be a subject of additional future study.

IV. Conclusion and Future Work

It is increasingly critical to utilize methods proven to benefit learners in online learning environments. In this paper, we used the same Macroeconomics course to do four doer effect analyses: replicating the correlational and causal models of Koedinger et al. [11] using both courseware unit tests and final exam scores. By studying the doer effect in this comprehensive manner on a single course, we can confirm correlational and causal findings using different learning outcomes and directly compare results from an assessment that serves as a near transfer of learning with an assessment that serves as an intermediate/far transfer of learning.

Our research question, "Can the correlational doer effect be replicated using both courseware unit tests and final exam scores?" was affirmed. Both linear regression models found that doing was positively significant, while reading was either not significant (unit tests) or negatively significant (final exam). Therefore, the doer effect ratio for both models was ∞.

Confirming the doer effect correlational analysis, it was also reasonable to ask our second research question: "Can the doer effect be accounted for by student characteristics?" Koedinger et al. [10] also checked to ensure that usage of the OLI courseware was still significant when accounting for student characteristics and found that it was. In this work, using a linear regression model that controlled for student characteristics—gender, underrepresented status, first-generation status, Pell eligible status, and age—we found that doing was still significant. Doing formative practice is still more effective than reading no matter student demographics.

Our research question—"Can the causal doer effect be replicated using both courseware unit tests and final exam scores?"—was positively answered. The courseware unit tests and final exam data produced results consistent with those of the original study. Replicating the findings of Koedinger et al. [11] using courseware designed with the same learning science principles but in a different domain and at a different higher education institution extends the generalizable nature of the doer effect findings. By engaging with a learning by doing design—formative practice questions integrated into the learning material—students activate the doer effect and increase their learning gains. This analysis confirms that even when controlling for an outside variable, doing the formative practice within the courseware caused better performance on both in-course unit tests and an external final exam. Doing practice *causes* better learning.

Some interesting observations can be drawn by comparing results across the different regression models. The standardized doing coefficient was always positive and highly significant, whereas the standardized reading coefficient was not always significant or positive. The doing coefficient also was always much larger than the reading coefficient. Although reading was positively significant and thus allowed computation of a numerical doer effect in only one of the six regressions performed in this work, we can still compare the magnitudes of the doing coefficients themselves across studies.

The doing coefficient was larger for the unit tests than the final exam in the correlational analyses, 0.4472 vs. 0.3655.

This was also the case for the within-doing coefficients in the causal analyses, 0.1629 vs. 0.1146. Although only a single course was studied, this is qualitatively consistent with a priori expectations; the unit tests are more proximal to the formative practice and hence less likely to be affected by learning decay, and alignment with the practice would be expected to be better with in-course summative assessments than an independently developed final exam.

Not only was it found that the doer effect could not be accounted for by student characteristics, inclusion of student characteristics as covariates had minimal impact on the value of the doing coefficient. Controlling for student characteristics changed the doing coefficient by -2.6 % and -5.7% in the unit test and final exam correlational models, respectively.

This work is to our knowledge the first to compare correlational and causal doer effect models on the same course. While not possible to make generalizable observations from a single course, especially quantitatively, it is interesting to note that significance of doing in the correlational model corresponded to significance in the causal model for both the unit tests and final exam. Should it turn out that observing a correlational doer effect generally tends to go along with a causal doer effect, this may be of practical interest because the data needed for a correlational study is simpler and available much more often than the data needed for a causal study. All these trends will be the subject of future investigation planned using a larger sample of courses.

The data available through courseware enable analysis and evaluation of learning principles, such as this one. Through large-scale data collected in a natural learning environment, learning analytics can broaden support for learning science concepts and strategies and provide generalizable results for additional learning contexts. In this particular case, the Macroeconomics courseware provided a comprehensive learning environment for students, but the final exam was what determined the course grade and final student outcome. This use-case may be similar to other higher education institutions where a high-stakes course assessment would take place as a proctored event outside of the learning environment. Identifying the doer effect using a final exam is encouraging because the potential for learning decay is greater than on a more proximal assessment, such as a unit test. What's more, separate development of the learning content and formative practice from the final exam could have made the doer effect more difficult to identify, but that was not the case. The use of a final exam for analysis may also be more typical of a college course where the content and exam are from different authors.

Learning engineering will continue to require not only collaboration of organizations and team members to engage in the LEP, but also the combination of different data sources to investigate learning principles in applied contexts. This study highlights the value of combining data from institutions and educational technology that collects large volumes of raw student data. Analysis for causality required both engagement data from the formative practice in the courseware as well as student learning outcomes from a final exam. As more data become available, combining data from different sources can accomplish valuable analysis of learning methods and principles. The doer effect research was critical to the design of the courseware environment during the LEP, and this process is furthered by sharing this replication research.

The significance of causal doer effect findings suggests at least two main avenues for future work. The first is to bring the learning by doing method to learning environments at scale, to provide as many students as possible with the learning benefits possible through the doer effect [20]. Doing causes learning, and these findings have been replicated in a variety of subject domains, using learning resources created by different organizations, and implemented at different institutions. The second goal of future work is to use these findings for iterative improvement in the LEP by identifying ways of increasing the amount of practice students do. While variation in the amount of practice students did in the progression of the course was necessary for the statistical models, it would be ideal if every student did effectively all the formative practice available. If doing causes learning, students should engage in as much formative practice as possible to leverage the causal doer effect and maximize its contribution to their learning outcomes. Future work can focus on the role of instructor implementation practice [22] and student motivation in increasing engagement.

ACKNOWLEDGMENT

We gratefully acknowledge Bill Jerome and Ken Koedinger for helpful discussions of this work. We also thank Margaret Hsaio for assisting in the preparation for this project.

REFERENCES

[1] R. Van Campenhout, B. G. Johnson, and J. A. Olsen, The Doer Effect: Replicating Findings that Doing Causes Learning. Presented at eLmL 2021 : The Thirteenth International Conference on Mobile, Hybrid, and On-line Learning. ISSN 2308-4367, pp. 1–6, 2021. Retrieved from: https://www.thinkmind.org/index.php?view=article&articleid=elml_2021_1_10_58001

[2] J. R. Anderson, A. T. Corbett, and F. Conrad, "Skill acquisition and the LISP tutor," Cognitive Science, vol. 13, pp. 467-506, 1989.

[3] L. W. Anderson et al. A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives (Complete edition). New York: Longman. (2001).

[4] P. Black, and D. William, "Inside the black box: raising standards through classroom assessment." Phi Delta Kappan, vol. 92(1), pp. 81–90, 2010. https://doi.org/10.1177/003172171009200119

[5] J. Dunlosky, K. Rawson, E. Marsh, M. Nathan, and D. Willingham, "Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology." Psychological Science in the Public Interest, vol. 14(1), pp. 4–58, 2013. https://doi.org/10.1177/1529100612453266

[6] J. Goodell, M. Lee, and J. Lis, "What we discovered at the roots of learning engineering." In IEEE ICICLE Proceedings of the 2019 Conference on Learning Engineering, Arlington, VA, May 2019.

[7] IEEE ICICLE. "What is Learning Engineering?" Retrieved 01/11/2021 from: https://sagroups.ieee.org/icicle/

[8] A. Kessler and Design SIG colleagues. Learning Engineering Process Strong Person, 2020. Retrieved 01/11/2021 from https://sagroups.ieee.org/icicle/learning-engineering-process/

[9] P. A. Kirschner, J. Sweller, and R. E. Clark, "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, vol. 41, pp. 75–86, 2006. http://doi:10.1207/s15326985ep4102_1

[10] K. Koedinger, J. Kim, J. Jia, E. McLaughlin, and N. Bier, "Learning is not a spectator sport: doing is better than watching for learning from a MOOC." In: Learning at Scale, pp. 111–120, 2015. Vancouver, Canada. http://dx.doi.org/10.1145/2724660.2724681

[11] K. Koedinger, E. McLaughlin, J. Jia, and N. Bier, "Is the doer effect a causal relationship? How can we tell and why it's important." Proceedings of the Sixth International Conference on Learning Analytics and Knowledge, LAK 2016, pp. 388-397. http://dx.doi.org/10.1145/2883851.2883957

[12] K. R. Koedinger, R. Scheines, and P. Schaldenbrand, "Is the doer effect robust across multiple data sets?" Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018, pp. 369–375.

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

[14] Olsen, J., & Shackelford, S. (2021). Intersectionality and Incremental Value: What combination(s) of student attributes lead to the most effective adaptations of the learning environment? In: Sottilare R., Schwarz J. (eds) Adaptive Instructional Systems. HCII 2021. LNCS, vol. 12792. Springer. pp. 577–591. https://doi.org/10.1007/978-3-030-77857-6_41

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

[16] M. Serra-Garcia, and U. Gneezy, "Nonreplicable publications are cited more than replicable ones," In Science Advances, vol. 7, pp. 1–7, 2021. http://doi.org/10.1126/sciadv.abd1705

[17] H. A. Simon, "The job of a college president," Educational Record, vol. 48, pp. 68-78, 1967.

[18] J. Sweller, "The worked example effect and human cognition," Learning and Instruction, vol. 16(2), pp. 165–169, 2006. https://doi.org/10.1016/j.learninstruc.2006.02.005

[19] R. Van Campenhout, "Learning engineering as an ethical framework: A case study of adaptive courseware," In: R. Sottilare, J. Schwarz (eds) Adaptive Instructional Systems, HCII 2021, Lecture Notes in Computer Science, vol 12792, pp. 105–119, 2021. Springer, Cham. https://doi.org/10.1007/978-3-030-77857-6_7

[20] R. Van Campenhout, J. S. Dittel, B. Jerome, and B. G. Johnson, "Transforming textbooks into learning by doing environments: an evaluation of textbook-based automatic question generation." In: Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education, 2021. Retrieved from: https://intextbooks.science.uu.nl/workshop2021/files/iTextbooks_2021_paper_6.pdf

[21] R. Van Campenhout, B. Jerome, and B. G. Johnson, "The impact of adaptive activities in Acrobatiq courseware: Investigating the efficacy of formative adaptive activities on learning estimates and summative assessment scores," In: R. Sottilare, J. Schwarz (eds) Adaptive Instructional Systems, HCII 2020, LNCS, vol. 12214, 2020. Springer. pp. 543–554. https://doi.org/10.1007/978-3-030-50788-6_40

[22] R. Van Campenhout and M. Kimball, "At the intersection of technology and teaching: The critical role of educators in implementing technology solutions. IICE 2021: The 6th IAFOR International Conference on Education. Retrieved from: https://papers.iafor.org/submission59028/