The Doer Effect: Replicating Findings that Doing Causes 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*—There is a dire need for replication research in the learning sciences, as methods put forth for increasing student learning should be unequivocally grounded in reproducible, reliable research. Learning science research is not only a critical input in the learning engineering process during the development of educational technology tools, such as courseware, but also as an output after student data have been analyzed to determine if the learning methods used were effective for students in their natural learning context. Furthermore, research that can provide causal evidence that a method of learning is effective for students should be reproduced—and the generality for its use expanded—so that methods that cause learning gains can be widely applied. One such method is the doer effect: the principle that students who engage with more practice have higher learning gains than those who only read expository text or watch video. This effect has been shown to be causal in prior research through statistical modeling using data mined from natural learning contexts. The goal of this paper is to replicate this research using a large-scale data set from courseware used at a major online university. The learning-by-doing data recorded by the courseware platform were combined with final exam data to replicate the statistical model of the causal doer effect study. Results from this analysis similarly point to a causal relationship between doing practice and learning outcomes. The implications of these doer effect results and future learning science research using large-scale data analytics will be discussed.**

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

I. INTRODUCTION

Advances in educational technology are increasingly beneficial to learning, yet increasingly complex in nature. Courseware is one such digital tool, which is designed to provide a comprehensive learning environment for students and real-time data insights to instructors [19]. The creation of tools such as courseware, however, is a daunting task to undertake. It is no small feat to imagine and define what courseware should be, but even more complex is the development process. The courseware has an authoring platform, a data architecture, a student interface, and instructor tools and dashboards, which require software engineers, product managers, and data scientists to develop. Similarly, creating the content that goes into courseware requires subject matter experts, instructional designers, media specialists, projects managers, etc. Learning engineering—an emerging discipline itself—provides a process for development and contextualization of the goals that helps synchronize often disparate teams and processes. Proposed by Herbert Simon [15] and fostered at Carnegie Mellon University [5], 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 Carnegie Mellon's Open Learning Initiative (OLI) to apply learning science and a student-centered approach to developing courseware [17].

Learning engineering as a practice supports learners and their development through the application of the learning sciences to human-centered engineering design methods and data-driven decision making [6]. 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 [7]—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.

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 passive reading or video watching the student does [10]. Studies have previously shown correlational support for this principle [9]. 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. [9] used data collected from students engaged with a MOOC course paired with courseware developed by OLI to investigate the doer effect. In their initial research, they found the learning effect of doing the formative practice was six times larger than that of reading. Follow-up analysis [10] [11] sought to determine whether this effect was causal. A statistical design involving within- and outside-unit doing, reading and watching (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 [10]: "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 [14]. 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. 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 [17]. 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) [10]. 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 [11]. Investigating an entirely different subject domain built independently—yet using the same learning science principles—strengthens the external validity of a causal relationship.

This study uses data from a business course, which is a domain outside of the STEM subjects originally analyzed, and a final exam to measure learning outcomes instead of unit tests. The final exam could potentially impact doer effect findings due to the increased learning decay that could occur over time when compared to unit tests.

Given the intention of this study to replicate causal doer effect findings, our research question is: Can causal doer effect findings be replicated on a final exam data set, generated from a competency-based online university course? To answer this, we will outline the required parallel features for this replication study in Section 2—from the learning by doing courseware environment, to the description of regression model and its inputs, to the data used for analysis. Section 3 will provide the formulas used for the analysis, the results, and a discussion on the meaning of the replication findings. Section 4 concludes the paper with remarks on the importance of these replication findings for the learning science methods used herein, the role of learning engineering and the LEP in continuing learning science research, and the implications of these findings for future research.

II. Methods

A. Learning by Doing in Courseware

In order for this replication research to be parallel with the original study, the learning resource needed to be similar in the learning by doing approach. Learning by doing as a term has been used to describe different kinds of learning engagement (and not all use or encourage the use of scaffolding or feedback [8]), 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 [3]. The formative practice questions integrated with the content essentially act as no-stakes practice testing, which increases learning gains and retention [4]. 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 [18]. 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 1). Immediate, targeted feedback was shown to reduce the time it took students to reach a desired outcome [1] [12], and feedback in practice testing outperforms no-feedback testing [4] [13]. Formative practice with targeted feedback provides scaffolding and examples that support cognitive structures for effective learning [8] [13] [16].

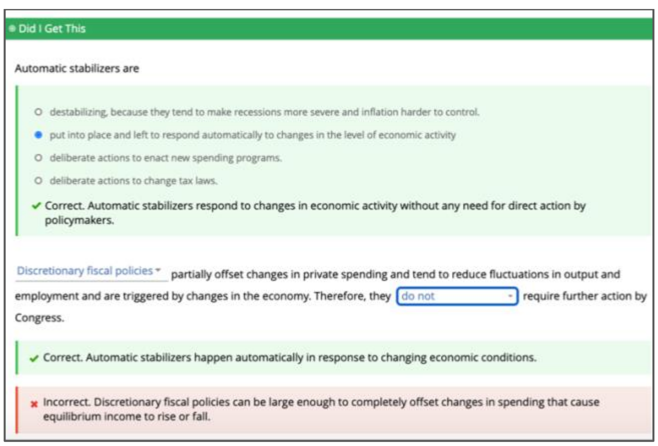

Figure 1. Formative practice questions from Macroeconomics

The courseware contains many features similar to those in used in the courses for the original study [10]. 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, 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 [2].

Western Governors University is an online, competency-based institution. 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. In addition to a unit on general learning strategies, there are six units of Macroeconomics content. Each unit contains an introduction, up to three modules of subtopic content, and a summary. Each module contains an adaptive activity and a quiz on the content from that module, and each unit summary contains a unit test cumulative to all modules in that unit [19].

Passing the WGU course depended solely on passing a 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 this study, the student's score on the first attempt at the final exam was used as the learning outcome.

B. The Model

A regression model developed by Koedinger et al. [10] 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.

The course analyzed in Koedinger et al. [10] 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.

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. An example competency developed for the Macroeconomics course is: ‘The Economic Way of Thinking - The graduate analyzes economic behavior by applying fundamental economic principles, including scarcity, opportunity cost, and supply and demand analysis.’

In order to apply the Koedinger et al. regression model [10] in the present study, 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. Henceforth, when referring to a *unit* of course content, 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.

C. The Data

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). 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 data set.

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. Following Koedinger et al. [10], the reading variables were defined as all visits to lesson pages where the student did not engage in any practice available on that page. There were 92,009 page visits for this group of students. The doing variables were defined as the number of formative practice opportunities a student attempted, including 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.

A total of 1,162 formative questions were included in the analysis, with 397,562 unique first attempts on these practice opportunities. 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. Unlike in the original study, watching was not investigated, as video was not a critical component of the courseware.

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.

III. RESULTS & DISCUSSION

For each of the 3,120 students in the data set, there is an observation (row) for each of the six competencies, bringing the total number of observations to 18,720. 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. [10], we use a mixed effects model to investigate the within-unit and outside-unit reading and doing relationships with learning outcomes. Reading, doing and competency score values were converted to *Z*-scores before regression to better enable comparison of the reading and doing effects. The R formula used to fit the model is below.

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

This shows that a linear mixed effects regression model was fit using the `lmer` function. The regression formula shows (normalized) competency score modeled as a function of within- and outside-unit reading and doing, with a random intercept per student and competency to address the lack of independence of the observations noted above.

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 following R code illustrates this test for the within-reading coefficient:

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

TABLE 1. DOER EFFECT REGRESSION ANALYSIS RESULTS.

Learning Method	*Location*	*Normalized Estimate*	*Std. Error*	*t-Value*	*Pr(>|t|)*
	(intercept)	0.0000	0.1256	0.000	1.0000
Doing	within-unit	0.1146	0.0099	11.613	< 2.2e-16 ***
	outside-unit	0.1556	0.0132	11.773	< 2.2e-16 ***
Reading	within-unit	-0.0125	0.0091	-1.367	0.1729
	outside-unit	-0.0604	0.0130	-4.645	3.432e-06 ***

The results of the regression analysis are presented in Table 1. 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.

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 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. The size of the doer effect, taken as the ratio of the standardized doing and reading coefficients, is also of interest. Previous work by Koedinger et al. [9] [10] found the effect of doing on outcomes was about six times greater than reading. In this study, however, we cannot compute a size for the doer effect because within-unit reading was not significant. Koedinger et al. [10] reported such cases as an effect ratio of ∞.

An interesting note is that the outside-unit reading coefficient was significant but negative, showing an overall negative relationship between the amount of outside-unit reading and final exam 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. Our research question—"Can causal doer effect findings be replicated on a final exam data set, generated from a competency-based online university course?"—was positively answered. The courseware and final exam data produced results consistent with those of the original study replicated. Replicating the findings of Koedinger et al. [10] 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 an external final exam. Doing practice *causes* better learning.

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

 ISBN: 978-1-61208-873-0

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 high-stakes assessment. 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 [18]. 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 this statistical model, 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 [20] 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] J. R. Anderson, A. T. Corbett, and F. Conrad, "Skill acquisition and the LISP tutor," Cognitive Science, vol. 13, pp. 467-506, 1989.

[2] 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).

[3] 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

[4] 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

[5] 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.

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

[7] 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/

[8] 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

[9] 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

[10] 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

[11] 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.

[12] 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

[13] 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

[14] 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

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

[16] 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

[17] 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, in press, 2021.

[18] 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 06/30/2021 from: https://intextbooks.science.uu.nl/workshop2021/files/iTextbooks_2021_paper_6.pdf

[19] 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

[20] 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 06/30/2021 from: https://papers.iafor.org/submission59028/

