Reducing Achievement Gaps: Exploring How Learning by Doing Can Support Diverse Student Success

Martha Hubertz & Patsy Moskal
University of Central Florida
Orlando, USA
email: martha.hubertz@ucf.edu, patsy.moskal@ucf.edu

Rachel Van Campenhout & Benny G. Johnson
VitalSource
Raleigh, USA
email: {rachel.vancampenhout, benny.johnson}@vitalsource.com

***Abstract*— Formative practice applied for a learning-by-doing approach is widely known to be an effective learning method for all students, but for disadvantaged students in particular. Different student populations—ethnic minorities, first generation college students, or economically disadvantaged—have historically had achievement gaps in higher education. Institutions have a responsibility to support diversity, equity, and inclusion, and utilize pedagogical practices and learning technology that can reduce the disparities in student success. This study analyzes a psychology course at the University of Central Florida that assigned formative practice in courseware to determine how this learning-by-doing method impacted exam scores for these student populations.**

Keywords-formative practice; doer effect; diversity; achievement gaps; student success.

I. INTRODUCTION

Higher education institutions have an obligation to provide equitable support for all students. As education serves as a powerful tool to address historical and systemic inequities, colleges and universities must enhance support mechanisms for student populations often at greater risk, such as racial or ethnic minorities, first-generation college students, and economically disadvantaged students. The University of Central Florida (UCF), one of the largest public universities in the United States, enrolls significant numbers of students from these groups, making it essential to evaluate how educational technology impacts their learning experiences and work to mitigate disparities in success.

Formative practice—low or no-stakes practice questions—has been found to be effective for all students, but for disadvantaged students most of all [1]. Research suggests that learning by doing could help mitigate disparities in student outcomes [2]. A meta-analysis found that active learning in STEM courses reduced the likelihood of student failure by 1.5 times compared to traditional instruction [3]. Theobald et al. [4] reviewed literature on active learning's impact on Black, Latino, Indigenous, and low-income students, concluding that active learning generally narrows achievement gaps in exam scores and improves passing rates.

Through the ability of digital learning platforms to collect extensive, high-quality micro-level data, we can gain valuable insights into learning processes for formative practice. For instance, data from courseware that integrates formative practice with text content through a learning-by-doing approach have highlighted the learning science principle known as the doer effect. Engaging in practice activities while reading has demonstrated an effect on learning approximately six times greater than reading alone, with studies confirming this relationship as causal [5][6][7]. Further analyses controlling for student characteristics, including minority status, gender, and age, found that the doer effect persists across diverse student groups [5][8].

Courseware with formative practice has been used at UCF in an online Psychology of Sex and Gender course since spring 2020 and prior research had found that assigning the formative practice increased student engagement and exam scores [9]. Given the high proportion of at-risk and disadvantaged students enrolled in the course, a post-hoc analysis was planned to investigate the relationship between learning by doing and learning outcomes for these students. This investigation required collaboration between the university and education technology company in order to combine data sources needed for this study. Notably, the courseware does not collect any student demographic data. Although its predictive models support adaptive activities and instructor dashboards, these models exclude demographic information for both legal and ethical considerations. In this case, the absence of a compelling need to incorporate demographic data guided this approach. There are strong arguments for setting boundaries to protect marginalized groups in machine learning applications, especially when that demographic data may not be necessary or appropriate [10]. Similarly, Baker [11] argues that demographic data in predictive analytics is both controversial and less actionable compared to learning behavior. Research further supports this view, indicating that learning data alone is a strong predictor of student success, outperforming other readiness assessments and demographic variables [12]. Consequently, investigating how this learning-by-doing environment supports specific student groups necessitates collaboration with the university that possesses the relevant demographic information. This investigation is guided by the primary research question: Can assigning learning-by-doing courseware help reduce achievement gaps among at-risk student populations in a psychology course at UCF?

 ISBN: 978-1-68558-271-5

In Section II, the technology, course context and implementation strategies, and data preparation are all described. In Section III, results are presented first using exploratory data analysis—including descriptive statistics and data visualizations—and second using the doer effect analysis and regression models to determine the significance of doing formative practice and student characteristics on exam scores. Section IV discusses limitations, conclusions, and future work.

II. METHODS

The courseware was generated using artificial intelligence and the volume of formative practice required for effective learning-by-doing from textbook materials [13]. These AI-generated questions underwent rigorous evaluation using student data, including data from UCF courses. Findings reveal that AI-generated questions perform comparably to human-authored ones, with students perceiving no significant differences, thereby validating their effectiveness in learning-by-doing environments [13].

This courseware served as the primary learning material and the instructor assigned the formative practice activities as completion-based homework. All sections of the course were synchronous online, mitigating some impacts from COVID-19 regarding modality during that time. In spring 2020 (S20), the assignment was worth 2% of the students' grade, whereas in spring 2021 (S21) and spring 2022 (S22) it was worth 20% of the grade. Prior research found these implementation changes resulted in increased student engagement and improved exam scores [9].

To assess the impact of the learning-by-doing method on various student populations, it was necessary to integrate multiple data sources. The first data source was raw clickstream data from the courseware platform, capturing timestamps for actions such as page visits and question interactions. This information is linked to anonymized numeric student identifiers, ensuring privacy. After obtaining institutional review board approval for this post-hoc analysis, these numeric identifiers were provided to UCF where grade data and student characteristics were added to a spreadsheet. Using the numeric identifiers allowed the VitalSource team to combine the student characteristics with the data set of millions of clickstream events for anonymous analysis.

While there were a combined 388 students across semesters at the start of the course, 81 students were removed for not completing the course (Grade = "W", "WD", "S" or "U")—a percentage not uncommon given the enrollment process prior to the add/drop date, and the community service hours required as a designated service-learning course. An additional 19 students were removed for not taking all 3 exams (18 were female, 12 were Hispanic). There were 287 students remaining in the data set for analysis: 90% female and 10% male; 50% white, 31% Hispanic/Latino, 10% Black/African American, 3% mixed race; 78% full time and 22% part time; 80% non-first generation and 20% first generation college students; 55% non-Pell eligible and 45% Pell eligible (Pell eligible being a proxy for economically disadvantaged). Note that as a post hoc analysis of natural learning contexts, student characteristics are a reflection of the course population and are not balanced, especially gender in this instance.

III. RESULTS

A. *Engagement*

The first step of investigating the data is to gain insight into how students engaged with the courseware, often related to course policies and implementation strategies. In Figure 1, each semester is shown as a graphic visualization where the number of students are on the y-axis and each page of the course is represented linearly on the x-axis. In this way, time is also approximated on the x-axis, as students move chronologically through the courseware over the course of the semester. For each page of the courseware (a vertical slice of the graph), the blue dot represents the number of students who did the reading, the red dot is the number of students who did the formative practice, and the green dot is the quiz. In the S20 graph, there is a vertical gap between the reading and doing dots, indicating some students were reading without doing. This reading-doing gap is fairly typical. However, in S21 and S22, the reading-doing gap is nonexistent. These semesters also show less attrition within units and across the course. The change in incentive for doing the formative practice had a large impact on student engagement patterns.

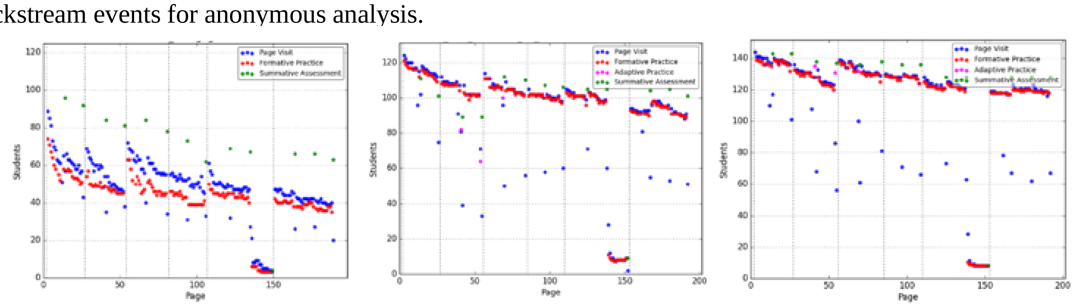

Figure 1. Engagement graphs for S20, S21, and S22 (left to right).

TABLE I. FORMATIVE PRACTICE COMPLETION BY SEMESTER

Semester	Students	Mean	STD	MIN	25%	50%	75%	MAX
Spring 2020	62	395.90	283.74	0.0	52.25	582.0	652.0	707.0
Spring 2021	99	610.43	149.95	0.0	618.50	665.0	672.5	888.0
Spring 2022	126	627.77	123.33	0.0	658.25	667.0	670.0	727.0

B. *Exam Scores*

The largest graded component of this course was the exam, which provides a quantitative measure of comparison across semesters and between demographic groups. Table II shows combined exam scores for each semester. As would be hoped from the increase in student engagement with the formative practice, exam scores increase across the 25th, 50th, and 75th percentiles. S22 has a very low scoring minimum outlier which does impact the overall mean, but the trend is as expected.

Viewing exam scores by gender (Table III), a few trends emerge, though the smaller proportion of males compared to females is important to keep in mind. For both S20 and S21, males had higher exam scores at the 25th and 50th percentile, but not the 75th percentile. For S22, females had higher scores for the 25th, 50th, and 75th percentile.

Viewing exam scores by ethnicity (Table IV) shows trends across semesters as well. With a few exceptions (S20 25th and S21 50th), white students had the highest exam scores across percentiles. Despite having the second highest population represented, Hispanic/Latino students typically had the lowest scores, with the exception of S22, where they surpassed the Black/African American group at the 25th and 50th percentile.

Table V reviews exam scores by Pell eligibility status, often used as a proxy for economic status. Students who were not Pell eligible outperformed those who were across each percentile and semester. These groups were most closely aligned in performance in S21, also where the proportion of students who were and were not eligible were nearly equal. Table VI shows students who were full time outperformed those who were part time across all percentiles and semesters (except S21 25th). Table VII shows that students who were first generation college students performed worse than their peers across all semesters.

Examining these exam scores by different student groups clearly shows the achievement gap reported in research. However, further analysis will determine if these differences are significant.

TABLE II. STUDENT EXAM SCORES BY SEMESTER

	Students	Mean	STD	MIN	25%	50%	75%	MAX
Spring 2020	62	450.23	68.88	282.0	405.00	466.5	494.25	594.0
Spring 2021	99	469.14	69.67	307.0	410.50	472.0	527.50	595.0
Spring 2022	126	467.66	78.78	183.0	420.75	477.0	528.00	609.0

TABLE III. EXAM SCORES BY GENDER FOR ALL SEMESTERS

Semester	Gender	Students	Mean	STD	MIN	25%	50%	75%	MAX
Spring 2020	Female	59	448.83	70.30	282.0	399.00	462.0	495.00	594.0
	Male	3	477.67	14.43	461.0	473.50	486.0	486.00	486.0
Spring 2021	Female	86	468.67	71.68	307.0	409.75	466.0	533.50	595.0
	Male	13	472.23	56.87	370.0	451.00	478.0	511.00	547.0
Spring 2022	Female	112	468.69	79.51	183.0	425.25	477.0	531.00	609.0
	Male	14	459.43	74.90	348.0	397.50	464.0	518.25	573.0

TABLE IV. EXAM SCORES BY RACE FOR ALL SEMESTERS

Semester	Ethnicity	N	Mean	STD	MIN	25%	50%	75%	MAX
Spring 2020	White	31	464.32	71.59	282.0	417.00	486.0	505.50	594.0
	Hispanic/Latino	21	426.95	63.50	315.0	372.00	426.0	486.00	501.0
	Black/African American	7	454.29	70.40	330.0	433.50	450.0	487.50	558.0
Spring 2021	White	47	473.70	70.65	340.0	424.00	466.0	538.00	595.0
	Hispanic/Latino	34	456.56	73.77	307.0	406.00	457.0	525.25	595.0
	Black/African American	11	475.55	61.36	388.0	418.00	496.0	532.00	547.0
Spring 2022	White	66	478.56	72.69	309.0	438.75	484.5	531.00	609.0
	Hispanic/Latino	35	457.23	77.74	255.0	397.50	466.0	513.00	590.0
	Black/African American	12	429.50	112.23	183.0	371.25	442.5	518.25	552.0

TABLE V. EXAM SCORES BY PELL ELIGIBLE FOR ALL SEMESTERS

Semester	Pell	Students	Mean	STD	MIN	25%	50%	75%	MAX
Spring 2020	Yes	36	434.83	76.46	282.0	381.00	444.0	486.00	594.0
	No	26	471.54	50.76	360.0	432.75	481.5	502.50	561.0
Spring 2021	Yes	48	466.90	74.65	307.0	409.00	469.0	537.25	595.0
	No	51	471.25	65.32	343.0	424.00	472.0	524.50	595.0
Spring 2022	Yes	44	437.86	86.83	183.0	384.00	448.5	498.75	590.0
	No	82	483.65	69.52	309.0	450.00	493.5	533.75	609.0

TABLE VI. EXAM SCORES BY FULL TIME/PART TIME FOR ALL SEMESTERS

Semester	Course Load	Students	Mean	STD	MIN	25%	50%	75%	MAX
Spring 2020	Full Time	43	462.42	60.61	324.0	423.0	480.0	496.50	561.0
	Part Time	19	422.63	79.68	282.0	360.0	432.0	484.50	594.0
Spring 2021	Full Time	79	472.71	72.46	307.0	409.0	478.0	532.00	595.0
	Part Time	20	455.05	56.76	340.0	418.0	436.0	498.25	547.0
Spring 2022	Full Time	101	472.96	79.17	183.0	429.0	477.0	534.00	609.0
	Part Time	25	446.24	74.89	309.0	396.0	455.0	492.00	579.0

TABLE VII. EXAM SCORES BY FIRST GENERATION FOR ALL SEMESTERS

Semester	First Gen	Students	Mean	STD	MIN	25%	50%	75%	MAX
Spring 2020	Yes	15	435.80	65.75	324.0	390.00	444.0	483.00	558.0
	No	47	454.83	69.90	282.0	415.50	477.0	496.50	594.0
Spring 2021	Yes	21	460.71	78.10	307.0	403.00	463.0	523.00	571.0
	No	78	471.41	67.60	340.0	421.75	473.5	528.25	595.0
Spring 2022	Yes	22	436.86	82.23	300.0	362.25	439.5	490.50	585.0
	No	104	474.17	76.86	183.0	435.75	484.5	531.00	609.0

Figure 2 shows a visualization of students' total reading (x-axis) by total exam score (y-axis) for all three semesters. The scatterplot has a general triangular shape, nicknamed a data-tornado by the authors. A line fit to this data would likely have a slightly positive slope. No discernable difference in pattern is observed between semesters.

By contrast, Figure 3 shows a data wall; total exam scores and total doing has produced a nearly vertical plot. The formative practice was assigned and nearly all students did nearly all the practice, therefore, most dots are along the x-axis point for maximum assigned practice. Since the vertical line has a wide range of exam scores, does that mean the practice did not help improve scores? That is not possible to tell from this plot. If doing practice increased exam scores by 5%, we would still see the same range in scores. Interestingly, there are some dots to the right of the main line; those students did extra questions in the chapter that was not assigned and therefore have a higher practice total than their peers. Some students are to the left of the data wall, showing not all students did all practice. Notably, there are far more blue dots from the S20 semester to the left, which aligns with lower formative engagement.

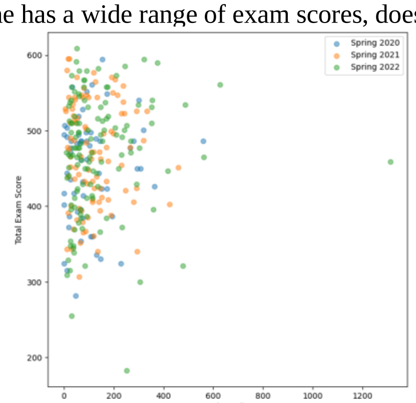

Figure 2. The data tornado: total reading by exam score.

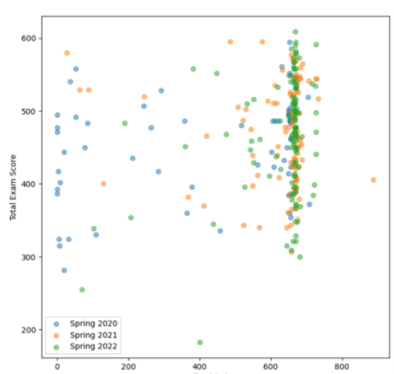

Figure 3. The data wall: total doing by exam score.

C. *The Doer Effect*

The doer effect analysis that is the foundational learning science principle supporting this learning by doing method requires data for reading, doing, and summative assessments—all of which we have for these courses. If we combined data from all semesters, we could find the same doer effect results; however, that is a misleading finding. Variation for within-student doing is necessary to determine the effects of doing practice on exam scores. There is variation in doing for S20, but very little variation in S21 or S22. In fact, the data wall is so vertical for S21 and S22, that the few outliers skew S21 positive and S22 negative in such a way that they cancel each other out, resulting in the misleading combined doer effect result. It is not possible to do a doer effect analysis for S21 and S22, but it is possible for S20 alone. In Table VIII, we see that the doing coefficient is significant, but the reading coefficient is not. The doer effect ratio (doing over reading coefficient) would be about 3, however, because reading is not statistically significant, the ratio is reported as infinity. This result is consistent with results reported by [6][7]. It is also likely that if the course had been closer to 100 students, reading would have become significant.

TABLE VIII. DOER EFFECT SPRING 2020

R1 (84)	Estimate Std.	Error	t Value	Pr(<|t|)
Intercept	461.196	10.238	45.047	<2e-16 ***
Total Reading	4.745	10.923	0.434	0.666
Total Doing	12.619	6.162	2.048	0.045 *

D. Regression Models

To determine how reading, doing practice, and student characteristics impacted exam scores, they were used as covariates in a linear regression model. In a linear regression for all semesters combined with all demographic covariates plus reading and doing (Table IX), the following covariates are significant: Hispanic/Latino, Pell eligible, full time/part time, and total doing. The linear regression for only S20 has the significant covariates of Hispanic/Latino, full time/part time, and total doing (Table X). This is similar to the results for all semesters combined. This semester produced the largest variation in doing practice so it is expected that doing would be significant to exam scores, as students who did more practice performed better than their peers.

TABLE IX. ALL SEMESTERS COMBINED

	Estimate Std.	Error	t Value	Pr(<|t|)
Intercept	509.092	20.048	25.394	<2e-16 ***
Male	5.745	14.492	0.396	0.692
American Indian/ Alaska Native	13.148	73.598	0.179	0.858
Asian	7.802	29.921	0.261	0.795
Black/African American	-11.531	15.088	-0.764	0.445
Hispanic/Latino	-22.072	10.143	-2.176	0.030 *
International	-14.078	36.406	-0.387	0.699
Multi-racial	0.355	24.760	0.014	0.989
Race Not Specified	5.890	71.622	0.082	0.935
Pell Eligible	-21.093	9.310	-2.266	0.024 *
First Generation	-6.812	11.421	-0.597	0.551
Age	-0.955	0.855	-1.118	0.265
Part Time	-22.155	10.928	-2.027	0.044 *
Total Reading	4.897	4.439	1.103	0.271
Total Doing	10.500	4.484	2.342	0.020 *

TABLE X. SPRING 2020

	Estimate Std.	Error	t Value	Pr(<|t|)
Intercept	463.784	38.068	12.183	<2e-16 ***
Male	61.713	42.155	1.464	0.150
Black/African American	1.490	28.579	0.052	0.959
Hispanic/Latino	-43.901	20.979	-2.093	0.042 *
Multi-racial	-6.142	67.090	-0.092	0.927
Pell Eligible	-23.869	18.131	-1.316	0.194
First Generation	-0.051	21.620	-0.002	0.998
Age	1.615	1.449	1.115	0.270
Part Time	-42.747	19.820	-2.157	0.036 *
Total Reading	5.282	10.847	0.487	0.629
Total Doing	14.026	6.586	2.130	0.038 *

In spring 2021, the results of the regression model in Table XI show there are no significant covariates. The lack of significance is overall positive, as the differences between demographic groups did not produce significant differences in exam scores. But does the lack of significance for doing mean it was not important for exam scores? No. If we recall the data wall, nearly all students did nearly all practice, and therefore there was not enough variation in doing to be statistically significant. In the context of implementation practices causing very high engagement, no significance indicates a successful engagement strategy.

TABLE XI. SPRING 2021

	Estimate Std.	Error	t Value	Pr(<|t|)
Intercept	529.536	51.570	10.268	<2e-16 ***
Male	9.838	22.975	0.428	0.670
American Indian/ Alaska Native	9.261	78.572	0.118	0.906
Black/African American	11.767	25.671	0.458	0.648
Hispanic/Latino	-17.474	17.032	-1.026	0.308
Multi-racial	20.231	32.025	0.632	0.529
Pell Eligible	2.309	15.737	0.147	0.884
First Generation	-11.475	19.242	-0.596	0.553
Age	-2.422	2.397	-1.010	0.315
Part Time	-12.214	20.458	-0.597	0.552
Total Reading	2.659	11.450	0.232	0.817
Total Doing	-10.011	11.010	-0.909	0.366

In spring 2022, the regression model in Table XII shows significant covariates of Pell eligible, total doing, and marginal significance for Black/African American. The exam scores had a wider distribution for S22 than for S21, so finding significant covariates is not unexpected. The doing covariate being significant again is indicative of a wider variation of doing for some students that did correlate to exam scores. The S22 semester did have some extreme outliers for exam scores that could also be contributing to the significance results.

TABLE XII. SPRING 2022

	Estimate Std.	Error	t Value	Pr(<|t|)
Intercept	522.164	29.993	17.410	<2e-16 ***
Male	9.799	22.230	0.441	0.660
Asian	3.995	31.990	0.125	0.901
Black/African American	-41.823	24.853	-1.683	0.095 .
Hispanic/Latino	-19.644	16.390	-1.199	0.233
International	-10.596	38.534	-0.275	0.784
Multi-racial	-32.696	54.366	-0.601	0.549
Race Not Specified	-0.526	74.509	-0.007	0.994
Pell Eligible	-34.953	15.779	-2.215	0.029 *
First Generation	-5.977	19.465	-0.307	0.759
Age	-1.723	1.236	-1.394	0.166
Part Time	-11.009	18.027	-0.611	0.543
Total Reading	8.164	5.715	1.429	0.156
Total Doing	29.253	11.951	2.448	0.016 *

IV. CONCLUSION AND FUTURE WORK

This study set out to explore whether learning-by-doing courseware can help reduce achievement gaps for at-risk student populations in a psychology course at the University of Central Florida. At its core, the research is motivated by the imperative for higher education institutions to provide equitable support mechanisms for all students, particularly those who have historically faced systemic barriers to academic success. By leveraging the rich behavioral data from formative practice embedded in digital courseware combined with exam scores and demographic data—we aimed to understand how such tools can contribute to a more inclusive and effective learning environment. The significance of this work lies in its potential to inform scalable, ethical interventions that support academic equity.

These results, while not proving a causal relationship, combine to provide a conclusion that the formative practice assigned in the courseware benefited all students. The variation in engagement for S20 gave a unique opportunity to do a doer effect regression analysis that gave correlational results in line with prior doer effect findings [6][7]. In all cases where a correlational doer was found (even in cases of infinity due to the reading covariate not being significant), a causal doer effect analysis was also confirmed [7], providing reasonable argument to expect the same would be found here if the conditions allowed for the causal analysis.

The relationship between the course policy of assigning practice, increased student engagement, and the impact on demographic disparities on exam scores is also supported by these results. By increasing the percentage of the students' grade for doing the formative practice, in S21 student engagement increased to the point that doing no longer became significant in the linear regression model for exam scores. No other demographic characteristics were significant—a positive finding.

There are limitations to this analysis. Because the sample size in S20 was only 62 students, inclusion of several demographic categories limited statistical power. We acknowledge that a larger number of independent variables may overfit the data, but we included these variables to explore potential achievement gap trends. Comparing different cohorts of students always brings variation that cannot be controlled for. The differences between the S21 and S22 results could easily be the result of the constitution of those students' characteristics. Future research could study results excluding extreme outliers to investigate the impact of those students on overall results. Future research should also examine more semesters to better identify trends over time, including semesters prior to S20 where there was no formative practice available to provide a different control measure for comparison.

The demographic data reveals that there is an achievement gap for student populations related to race, first generation status, and economic status. As education is an essential component for student success later in life, supporting student success with a focus on reducing or eliminating the achievement gap for these groups continues to be a vital mission in higher education. Any learning tool and pedagogical strategy that can work towards mitigating these achievement gaps should be embraced.

REFERENCES

[1] P. Black and D. Wiliam, "Inside the Black Box: Raising Standards Through Classroom Assessment," *Phi Delta Kappan*, vol. 92, no. 1, 2010, pp. 81–90. https://doi.org/10.1177/003172171009200119.

[2] L. DaVinci, "The Impact of Digital Learning on Minoritized and Poverty-Affected College Students: A Literature Review," *Every Learner Everywhere*, 2023. Retrieved: April, 2025, https://www.everylearnereverywhere.org/resources/the-impact-of-digital-learning-on-minoritized-and-poverty-affected-college-students-a-literature-review/

[3] S. Freeman et al., "Active Learning Increases Student Performance in Science, Engineering, and Mathematics," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, 2014, pp. 8410–8415. https://doi.org/10.1073/pnas.1319030111.

[4] E. J. Theobald et al., "Active Learning Narrows Achievement Gaps for Underrepresented Students in Undergraduate Science, Technology, Engineering, and Math," *Proceedings of the National Academy of Sciences*, vol. 117, no. 12, 2020, pp. 6476–6483. https://doi.org/10.1073/pnas.1916903117.

[5] 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," *Proceedings of the Second ACM Conference on Learning@Scale*, Vancouver, BC, Canada, 2015, http://dx.doi.org/10.1145/2724660.2724681.

[6] 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 & Knowledge*, Edinburgh, United Kingdom, 2016, http://dx.doi.org/10.1145/2883851.2883957.

[7] R. Van Campenhout, B. Jerome, B. G. Johnson, "The Doer Effect at Scale: Investigating Correlation and Causation Across Seven Courses," *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023)*, 2023, https://doi.org/10.1145/3576050.3576103.

[8] R. Van Campenhout, B. G. Johnson, and J. A. Olsen, "The Doer Effect: Replication and Comparison of Correlational and Causal Analyses of Learning," *International Journal on Advances in Systems and Measurements*, vol. 15, no. 1&2, 2022, pp. 48–59.

[9] M. Hubertz and R. Van Campenhout, "Leveraging Learning by Doing in Online Psychology Courses: Replicating Engagement and Outcomes," *The Fifteenth International Conference on Mobile, Hybrid, and On-line Learning*, 2023, pp. 46–49. Retrieved: April, 2025. https://www.thinkmind.org/index.php?view=article&articleid=elml_2023_2_60_50025.

[10] E. Mayfield et al., "Equity Beyond Bias in Language Technologies for Education," *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2019, pp. 444–460. https://doi.org/10.18653/v1/W19-4446.

[11] R. S. Baker, "Stupid Tutoring Systems, Intelligent Humans," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 2, 2016, pp. 600–614. https://doi.org/10.1007/s40593-016-0105-0.

[12] J. Olsen and S. Shackelford, "Intersectionality and Incremental Value: What Combination(s) of Student Attributes Lead to the Most Effective Adaptations of the Learning Environment?" in *Adaptive Instructional Systems. HCII 2021*, R. Sottilare, J. Schwarz (Eds.), LNCS, vol. 12792, Springer, 2021, pp. 577–591. https://doi.org/10.1007/978-3-030-77857-6_41.

[13] 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," *Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education, CEUR Workshop Proceedings*, 2021. Retrieved: April, 2025. http://ceur-ws.org/Vol-2895/paper06.pdf.