ADOPTING DATA SCIENCE CURRICULA: A STUDENT CENTRIC EVALUATION

S. Wang[1], V. P. Janeja[2], D. Harding[3], C. von Vacano[3], D. Lobo[3]

[1]Mills College (UNITED STATES)
[2]University of Maryland, Baltimore County (UNITED STATES)
[3]University of California, Berkeley (UNITED STATES)

Abstract

With the advent of data science as a new discipline and high demand for a skilled workforce, educators are increasingly recognizing the value of translating courses and programs that have been shown to be successful and sharing lessons learned in increasing diversity in data science education. In this paper, we describe and analyze our experiences translating a lower-division data science curriculum from the University of California, Berkeley's Data 8 Foundations of Data Science course to other settings with very different student populations and institutional contexts at the University of Maryland, Baltimore County and Mills College during Fall 2021. It is essential to motivate students to meaningfully take part in their journey into a data science career. We wanted to consider their perceptions and motivations to take the foundations course and the next steps emerging from the foundations course.

We evaluated how students were receiving the course and curriculum with the adaptations across the three institutions. Two identical detailed surveys were administered, once at the beginning of the semester and then at the end, to study the impact of the course and its variations on the students. The survey results showed gains in students' motivations and perceptions of being in data science across all three institutions. Our findings emphasized the importance of adapting courses and programs to existing curricula, student populations, cyberinfrastructure, and faculty and staff resources in the context of the institutions. Such adaptation can help students develop their understanding of data science career pathways and help hone their motivations, which can lead to a more engaged workforce supporting data science careers.

Keywords: Data science, adoption, student perspective.

1 INTRODUCTION

Data science has emerged as a new discipline and as such it poses an important challenge for undergraduate education to meet the needs of a burgeoning and yet diverse workforce [11]. Educators are increasingly recognizing the value of adapting and translating courses and programs that have been shown to be successful and sharing lessons learned in increasing diversity in data science education. Yet pedagogical practices and curricula cannot simply be copied from one setting to another. Instead, faculty and administrators wishing to adopt innovations from other institutions must make systematic and thoughtful efforts to translate what has worked in one setting into their own, fitting it into existing curricula and tailoring it to their own student populations' interests and needs.

Many data science curricula have been proposed and implemented [4, 5, 7, 8, 9, 11, 12]. One early developer of a robust curriculum of data science education is the University of California, Berkeley (UC Berkeley) [4], which began its focus on scaling lower-division data science courses. The UC Berkeley model has already attracted national attention and interest in replication, with early indications suggesting that parts of the model can be successfully transferred to other institutions. A recent report from the National Academy of Sciences cites UC Berkeley's program as an exemplar in teaching and broadening participation in data science [11]. The approach has sparked widespread interest; indeed, around 400 faculty from higher education institutions across the country attended a UC Berkeley workshop on data science pedagogy in the summer of 2022 [3], with more than 15 sessions focusing on UC Berkeley's model.

In this paper, we describe and analyze our experiences translating a lower-division data science curriculum from the University of California, Berkeley's Data 8 Foundations of Data Science course (Foundations) to other settings with very different student populations and institutional contexts at the University of Maryland, Baltimore County (UMBC) and Mills College during Fall 2021. We evaluate student experiences across the three institutions, in terms of their motivations and perceptions about

data science and their sense of belonging in data science, in a comparison of student responses before and after taking the Foundations course.

We begin with the institutional contexts of UC Berkeley, the University of Maryland, Baltimore County and Mills College. We then discuss the ways in which the University of Maryland, Baltimore County, and Mills College implemented the data science Foundations course. Next, we present a summary of survey results to evaluate the impact of the Foundations course on student perceptions in a student cohort from 2021 across the three institutions. Finally, we discuss some of the findings that reflect our experiences in the adaptations and implementation.

2 INSTITUTIONAL CONTEXT

2.1 UC Berkeley

UC Berkeley is one of California's flagship public research universities with a student population of over 42,000 and over 350 different degree programs. Twenty-two percent of undergraduates are underrepresented minorities, 21 percent of undergraduates are transfer students, and 23 percent of first-year admits are first-generation students. We estimate that almost one-quarter of Berkeley's undergraduates take the Foundations course at some point in their college career. Despite its creation only four years ago, Data Science is now UC Berkeley's third-largest major (behind Computer Science and Economics).

2.2 UMBC

UMBC is well known in the space of inclusive excellence and prides itself on the diversity of the campus emphasizing student success. With well-established and respected programs, UMBC has a big impact in the region and across the global social and economic spectrum. The university has an emphasis on not only excelling in pedagogy but research as well. UMBC has over 13,000 students, and with 52% minority enrollment, it is a Minority Serving Institute (MSI). UMBC currently has the R1 Carnegie classification. The Information Systems department offered the Foundations course and is one of the largest departments at UMBC, offering several data science classes and related undergraduate specializations in cyber informatics and business analytics, and graduate tracks and certificates in data science and artificial intelligence.

2.3 Mills College

Mills College is a small, nationally renowned liberal arts women's college located in the San Francisco Bay Area with a student population of 960. As one of the most diverse liberal arts colleges in the country with 65% students of color, a Hispanic Serving Institution (HSI), 44% first-generation, and 58% LGBTQ+, Mills College has a strong record of academic success with its students and a deep commitment to equity, inclusion, and social justice. The Mills experience is distinguished by small, interactive classes, one-on-one attention from exceptional faculty, a culture of creative experimentation, and cutting-edge interdisciplinary learning opportunities which empower students to make a statement in their careers and communities.

2.4 Comparative View of the Three Institutions

The three institutions bring diverse perspectives in terms of the size of the institution, size of the Foundations class, and minority serving capacities, as shown in Table 1.

Table 1. Comparative View of the Three Institutions

Institution	***UC Berkeley***	***UMBC***	***Mills***
Carnegie classification	R1	R1	Liberal arts
Public/Private	Public	Public	Private
Total undergrad student body	31,800	10,835	660
Average Foundations class size	1500	18	20
Minority serving status	[no]	MSI	HSI
URM population	22%	52%	65%
First-gen population	23%	25%	44%
Female identifying undergrads	54%	44%	100%

For example, UMBC and Mills both bring in a high level of diversity in terms of minority students and first-gen students. However, the Foundations class sizes are much smaller at these two institutions compared to UC Berkeley, perhaps a reflection of the size of the student body, which is much larger at UC Berkeley. These variations bring unique challenges and opportunities in transferring the lessons learned from UC Berkeley to UMBC and Mills and vice versa.

3 IMPLEMENTATION AND ADAPTATIONS

Data 8 Foundations of Data Science course [2] offered at UC Berkeley is an entry-level, four-unit course designed for any major with no prerequisites. It teaches critical concepts and skills in computer programming and statistical inference, in conjunction with hands-on analysis of real-world datasets, including economic data, document collections, geographical data, and social networks. The format consists of three one-hour lectures and a two-hour lab each week.

UMBC and Mills adopted the Foundations course independently. UMBC adopted it as an IS department course (IS 296: Foundations of Data Science) and has offered it four times since the fall of 2020. Mills offered the Foundations course as a Data Science course (DATA80A: Data Science for Everyone) in the fall of 2021. Though implemented separately, there are many similar themes in the adaptations by the two institutions.

Adapting to institutional context: The UMBC adaptation took into consideration (a) the student body and their backgrounds, (b) the number of credits/hours they can dedicate to studying per week with additional workload that they might have to balance.

UMBC had to lower the number of credits to three (as compared to four at UC Berkeley) with limited contact hours to fit the course into existing frameworks and include it on a pathway to existing degree requirements. It was also important that the students did not have to pay for an additional course and increase time to graduation. As a result, UMBC adaptation made this course fulfill a programming requirement for one of the majors [13] and also became part of the X+CS effort [10].

Mills reduced the number of lab hours to one (versus two at UC Berkeley) in order to attract students to take the course and keep the number of contact hours in sync with the number of credits. The course also fulfilled a programming requirement for the general BS degree.

Adapting course infrastructure: In general, setting up the JupyterHub infrastructure was non-trivial and needed ramp-up time. For both institutions, the first run-through had challenges with versioning and interdependencies of Python libraries as well as security and authentication issues. However, once the infrastructure was set up and running, the students' experiences with authentication, pulling files from GitHub, accessing files, and working in Jupyter notebooks, for the most part, were smooth, easy, and seamless, which was the reason for setting up the JupyterHub in the first place.

UMBC also explored the use of Google Colab. The trick to having a strong working infrastructure was getting the right expertise at the right time, and the right amount of time from experts, which was really based on a partnership with the UMBC office of information technology and lessons learned from the Foundations infrastructure. With the heavy reliance on Jupyter Notebooks for this Foundations course, it was imperative to have this established well in advance with sufficient testing in place.

Mills had assistance from the IT staff at UC Berkeley and support from 2i2c [1] in setting up the JupyterHub. Mills personnel had experience working in Colab in the past in upper-division CS courses. However, Colab would not be suitable for novices working with Jupyter notebooks for the first time due to the multiple steps required to work with data files and images in Colab.

With regard to autograding, UMBC did not use it but instead had teaching assistants and faculty grade the assignments and projects, some of which were more flexible and needed individualized grading. Mills used autograding for labs (only) in order to provide immediate feedback. Manual grading for assignments and projects was possible as the class sizes were much smaller at both institutions than at UC Berkeley.

Adapting the content structuring: Both institutions, as a base, started with the same lab and project structures as the original Foundations course. A number of modifications were made to the content. For UMBC, around the halfway point the assignments and projects started becoming more flexible. UMBC adaptation added new homeworks, two different projects, and an ethics module. The overall aim in switching some of these was to include flexibility for students to bring in their own lived

experiences. These modifications included the following: (a) One of the homeworks the students did was an exploration of data science case studies to understand what was the type of research question being addressed with data, what worked or did not work; (b) Another homework was a lab (on linear regression) from UC Berkeley but students were asked to switch the dataset to something they were interested in exploring or ones they encountered in their case studies; (c) One of the UC Berkeley projects was replaced to include a tool exploration project where they could pick any data science tool (such as Weka, Rapid Miner, Google Cloud etc.) and explore it with the datasets they had encountered in their case study homework; (d) Another project was replaced, building on these homeworks, where students investigated a research question with their own or real-world data they encountered in the homeworks and had explored with the tools; (e) The module on ethics [14], including a guest lecture, and a lecture on ethical data life cycle was included midway through the semester.

Mills conducted the class in an interactive style, with time for discussions and Q&As. The topic of ethics was added in the middle of the semester with a new assignment, lab, and a guest lecture. The Mills adaptation adjusted and spread out the content of the course over the semester to accommodate the shortened lab hour and the time allocated for discussions, and consequently covered one less topic – classification. Mills replaced the third project with one focused on linear regression where students could use a prepared dataset or choose an appropriate dataset of their interest.

4 SURVEY RESULTS

Through the course offerings, we also wanted to evaluate how students were receiving the course and curriculum. Two identical surveys were administered at all three institutions, once at the beginning of the semester and then at the end. The survey consisted of over 60 questions including demographics, understanding the motivations for why students enrolled in the data science class, how they perceived data science, etc.

At UMBC and Mills, the class sizes for the Foundations course were small. UC Berkeley, on the other hand, had a very large cohort. To get a comparison of students with similar demographic backgrounds, we consider two sets of students from UC Berkeley, one as the full cohort and the other as a weighted subset corresponding to the demographic distributions of UMBC and Mills.

We selected a few relevant categories of questions to investigate to see if there was a change in students' perceptions and interests over the semester. It was important to assess what factors influenced students to take a data science course and their sense of belonging in the data science community. First, we focus on the motivations of students taking a data science class. Second, we focus on their perceptions of data science. We discuss the survey results across all three institutions in the cohort of students taking the Foundations class in 2021.

In the first set of questions, we considered why students took the data science class to understand their motivations. The questions were: Data science (DS) skills will improve my chances of getting a good job after graduation (Job); DS skills are important to my extracurricular activities (Activities); DS skills will help me make an impact and solve problems in society (Impact); DS is intellectually interesting to me (Intellectual); and Data literacy is important for everyone these days no matter what their career or major (Data literacy). These questions were scaled between 1 to 5 (from not at all important to extremely important).

Overall, UMBC and Mills both showed gains in students' motivations over the semester, as seen by comparing the mean scores for questions asked at the beginning of the semester versus the end (Figure 1). For example, UMBC and Mills students had a 0.36 point and 0.14 point improvement, respectively, in the combined mean score (the mean of mean scores) at the end of the semester when asked why they chose to be in data science. Interestingly at UMBC, students were motivated more by societal aspects, personal interests, and data literacy aspects than intellectual interests. At Mills, data literacy was not as important a motivation as compared to obtaining a job or extracurricular activities.

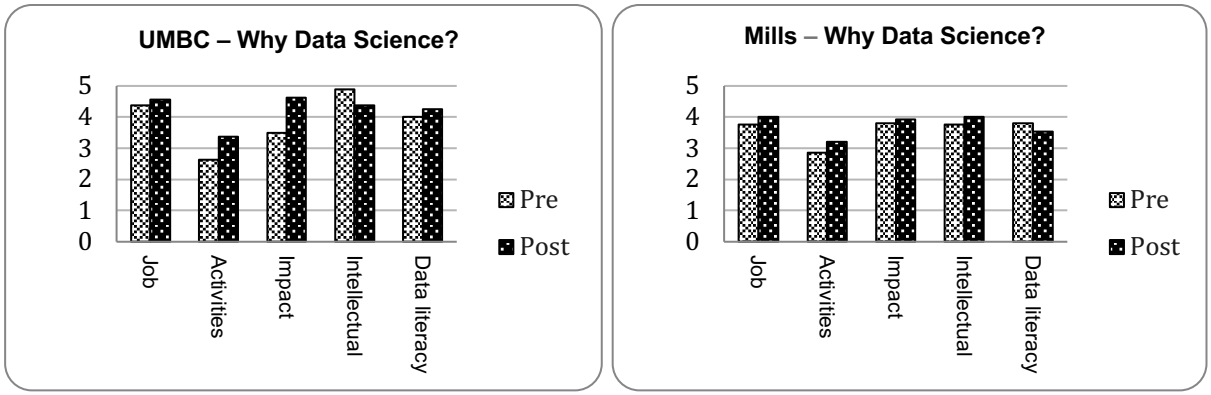

Figure 1. Motivations for Taking the Foundations Course for UMBC and Mills

Next, we considered student perceptions. The questions were: I see myself as a data science person; I see myself as a part of the data science community. These questions were scaled between 1 to 7 (from no, not at all to yes, very much).

Overall, UMBC and Mills students had a 0.82 point and 0.18 point improvement, respectively, in the combined mean score (the mean of mean scores) at the end of the semester when asked about their personal perceptions and belonging in data science (Figure 2). We see clear gains in student perceptions at the end of the semester at UMBC. We also see an improvement at Mills in the question about seeing oneself as a data scientist. We see a very minor dip in students seeing themselves as part of the data science community at Mills. This may be because at UMBC there are several community connection activities while Mills is in the process of developing these, such as the Data Science Scholars [15].

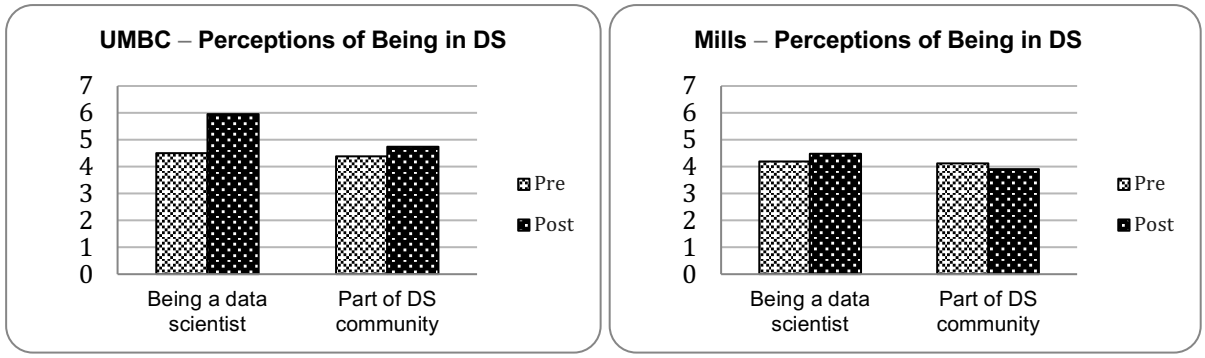

Figure 2. Perceptions of Being in Data Science for UMBC and Mills

UC Berkeley had a much larger cohort of students. In all questions about why students took data science at UC Berkeley, we see a very minor gain or, in some cases, a decline (Figure 3). This may be an effect of the large class sizes for the Foundations course. However, when we consider the perception questions, we see gains in mean scores in response to questions about students seeing themselves as data scientist and a part of the data science community.

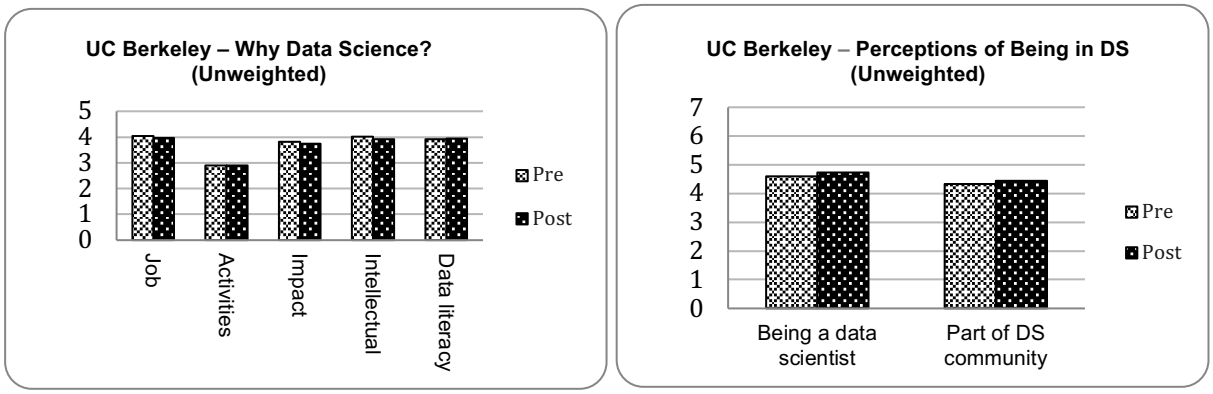

Figure 3. Motivations and Perceptions of Being in Data Science for UC Berkeley – Unweighted

To compare the results of the responses with distributions of students at UC Berkeley similar to those at UMBC and Mills, we also generated weighted samples, using propensity score weighting [6], from the UC Berkeley data with respect to race, gender, first-generation status, transfer status, and international student status (Figure 4). This subset of UC Berkeley's students shows a clear improvement across both sets of questions after taking the Foundations class.

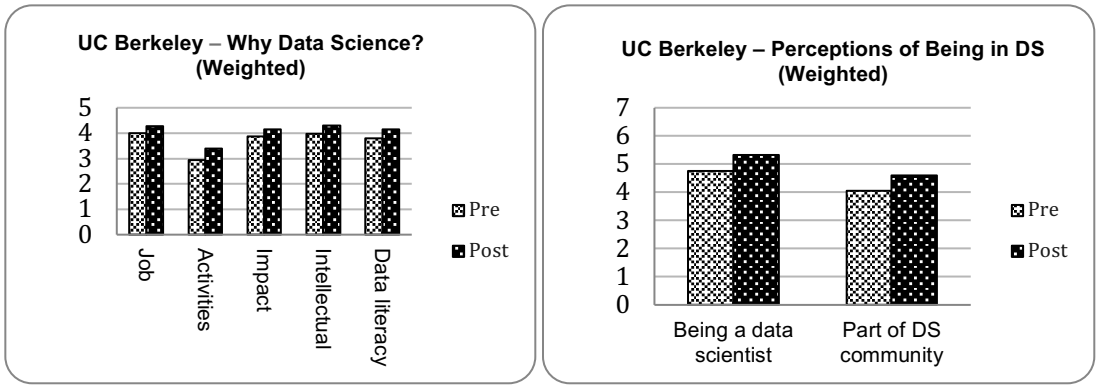

Figure 4. Motivations and Perceptions of Being in Data Science for UC Berkeley –Weighted

5 CONCLUSIONS

In this paper, we have described and evaluated our efforts to adapt and translate a lower-division data science course from a large public R1 university (UC Berkeley) to two different institutional contexts, an R1 university serving a diverse population (UMBC) and a liberal arts college for female-identified students (Mills College). Through detailed student surveys conducted at the beginning and the end of the semester, we saw gains in students' motivations and perceptions of being in data science across all three institutions. Our findings highlight the importance of beginning small before attempting to scale translated programs and the need to adapt courses and programs to existing curricula, student populations, cyberinfrastructure, and faculty and staff resources. Smaller class sizes open the possibility of more individualized assignments tailored to the majors, career interests, and social change motivations of students who may bring in lived experiences. While students across institutional contexts may need varying degrees of support, we found that often students from diverse backgrounds, if engaged deeply, show significant enthusiasm for data science and its applications.

ACKNOWLEDGEMENTS

This project is supported by National Science Foundation award #1915714. We thank Audrey Thomas and Jeff Royal for their assistance with survey implementation and Lynn U. Tran for survey development.

REFERENCES

[1] 2i2c, Interactive Computing for Your Community, 2022. Retrieved from https://2i2c.org/.

[2] *Foundations of Data Science*, UC Berkeley, 2015. Retrieved from http://data8.org/.

[3] *2022 National Workshop on Data Science Education*, UC Berkeley, 2022. Retrieved from https://data.berkeley.edu/2022workshop

[4] UC Berkeley, Berkeley Computing, Data Science, and Society Curriculum Overview, 2022. Retrieved from https://data.berkeley.edu/curriculum-overview.

[5] Ismail Bile Hassan, Thanaa Ghanem, David Jacobson, Simon Jin, Katherine Johnson, Dalia Sulieman, and Wei Wei, "Data Science Curriculum Design: A Case Study," in *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (Virtual Event, USA) (*SIGCSE '21*), Association for Computing Machinery, New York, NY, USA, pp. 529–534, 2021. Retrieved from https://doi.org/10.1145/3408877.3432443

[6] Jennie E. Brand and Yu Xie, “Who Benefits Most from College?: Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education,” in *American Sociological Review* 75, 2 pp. 273–302, 2010. https://doi.org/10.1177/0003122410363567arXiv:https://doi.org/10.1177/0003122410363567 PMID:20454549.

[7] Richard D De Veaux, Mahesh Agarwal, Maia Averett, Benjamin S Baumer, Andrew Bray, Thomas C Bressoud, Lance Bryant, Lei Z Cheng, Amanda Francis, Robert Gould, et al., “Curriculum guidelines for undergraduate programs in data science,” in *Annu Rev Stat Appl* 4 pp.15–30, 2017.

[8] Yuri Demchenko, Adam Belloum,Wouter Los, Tomasz Wiktorski, Andrea Manieri, Holger Brocks, Jana Becker, Dominic Heutelbeck, Matthias Hemmje, and Steve Brewer, “EDISON Data Science Framework: A Foundation for Building Data Science Profession for Research and Industry,” in *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pp. 620–626, 2016. https://doi.org/10.1109/CloudCom.2016.0107

[9] ACM Data Science Task Force, “Computing Competencies for Undergraduate Data Science Curricula,” Association for Computing Machinery, New York, NY, USA, 2021.

[10] S Mitchell, K Cole, and A Joshi. [n.d.], “X+CS: A Computing Pathway for Non-Computer Science Majors,” in *ASEE Mid Atlantic Section Spring Conference*, 2020 ([n. d.]). https://par.nsf.gov/biblio/10192230

[11] National Academies of Sciences Engineering, Medicine, et al., “Data Science for Undergraduates: Opportunities and Options”, 2018. https://doi.org/10.17226/25104

[12] Aimee Schwab-McCoy, Catherine M. Baker, and Rebecca E. Gasper, “Data Science in 2020: Computing, Curricula, and Challenges for the Next 10 Years,” in *Journal of Statistics and Data Science Education* 29, sup1, S40–S50, 2021. https://doi.org/10.1080/10691898.2020.1851159 arXiv:https://doi.org/10.1080/10691898.2020.1851159

[13] UMBC. [n.d.]. Bachelor of Arts in Business Technology Administration. https://informationsystems.umbc.edu/home/undergraduateprograms/undergraduate-degree-programs/bachelor-of-arts-in-businesstechnology-administration/

[14] Vandana P. Janeja, Maria Sanchez, “Rethinking Data Science Pedagogy with Embedded Ethical Considerations,” EDULEARN 2022, https://doi.org/10.21125/edulearn.2022.1964

[15] UMBC, Data Science Scholars, https://datasciencescholars.umbc.edu/