Fostering the Development of Earth Data Science Skills in a Diverse Community of Online Learners: A Case Study of the Earth Data Science Corps

Authors: Nathan A. Quarderer[1,2,3,4], Leah Wasser[5], Anne U. Gold[2,3,4], Patricia Montaño[6], Lauren Herwehe[4], Katherine Halama[1,3,4], Emily Biggane[7], Jessica Logan[7], David Parr[8], Sylvia Brady[8], James Sanovia[2,3,4,9], Charles Jason Tinant[10], Elisha Yellow Thunder[7,10,11], Justina White Eyes[10,12,13], LaShell Poor Bear/Bagola[10], Madison Phelps[10,11], Trey Orion Phelps[10], Brett Alberts[7,14], Michela Johnson[4], Naomi Jacquez[8], Kaiea Rohlehr[8], Emily Ward[2,3,4], Elsa Culler[1,2,3,4], R. Chelsea Nagy[1,2,3,4], Jennifer Balch[2,3,4]

1. Earth Lab
2. The Environmental Data Science Innovation and Inclusion Lab (ESIIL)
3. The Cooperative Institute for Research in Environmental Science (CIRES)
4. The University of Colorado Boulder
5. pyOpenSci
6. National Center for Atmospheric Research (NCAR)
7. United Tribes Technical College (UTTC)
8. Metropolitan State University of Denver (MSU)
9. The American Indian Higher Education Consortium (AIHEC)
10. Oglala Lakota College (OLC)
11. South Dakota State University (SDSU)
12. Native BioData Consortium
13. Dakota State University
14. United States Department of Agriculture (USDA)

Abstract

Today's data-driven world requires earth and environmental scientists to have skills at the intersection of domain and data science. These skills are imperative to harness information contained in a growing volume of complex data to solve the world's most pressing environmental challenges. Despite the importance of these skills, Earth and Environmental Data Science (EDS) training is not equally accessible, contributing to a lack of diversity in the field. This creates a critical need for EDS training opportunities designed specifically for underrepresented groups. In response, we designed the Earth Data Science Corps (EDSC) which couples a paid internship for

undergraduate students with faculty training to build capacity to teach and learn EDS using Python at smaller Minority Serving Institutions. EDSC participants are further empowered to teach these skills at their home institutions which scales the program beyond the training lead by our team. Using a Rasch modeling approach, we found that participating in the EDSC program had a significant impact on learners' comfort and confidence with technical and non-technical data science skills, as well as their science identity and sense of belonging in science, two critical aspects of recruiting and retaining members of underrepresented groups in STEM.

Background

Data are becoming increasingly abundant and diverse. From our smartphones to wearable computers to social media, we are constantly creating and being exposed to data. With this explosion of big data, there is a need for a data-literate workforce that is capable of accessing and analyzing these data in a way that is useful for a diversity of collaborators and decision makers (Henke et al., 2016). The Earth and Environmental Sciences (EES) specifically produces vast amounts of data, from satellite observations to in situ sensor measurements, while at the same time tackling challenging research questions that directly impact peoples' wellbeing, such as natural hazards (Iglesias et al, 2021) or impacts from climate change (Monteleoni, Schmidt & McQuade, 2013). At the nexus of big data and socially relevant questions, EES is experiencing an increasing need for well-trained researchers capable of working with large datasets in areas such as modeling weather and climate, managing large networks of environmental sensors, and analyzing the ever-growing quantities of information being streamed from satellites and other remote sensing devices (Gibert et al., 2018). This presents a need for

researchers with disciplinary backgrounds in EES who also possess the critical computational skills necessary to work with increasingly large datasets and conduct data-intensive environmental science investigations (Hampton et al, 2017). Minority Serving Institutions (MSIs) cater to a student population that often faces institutional barriers and other socioeconomic hurdles that can limit availability of the resources and educational opportunities needed to obtain and work with these data, leading to historically underrepresented communities who disproportionately lack access to and knowledge of these resources. This creates a need for Earth and Environmental Data Science (EDS) training opportunities (Wasser et al, 2022), with an emphasis on EES applications, designed for both learners and educators from underrepresented groups.

One approach to domain-specific, data analytic training in EES for members of historically marginalized communities is the Earth Data Science Corps (EDSC; Quarderer et al, 2023), led by the Earth Analytics Education Initiative at the University of Colorado Boulder (Earth Lab/The Cooperative Institute for Research in Environmental Sciences [CIRES]). This internship program partners with Tribal Colleges and Universities (TCUs) and a Hispanic Serving Institution (HSI) in the western U.S. to offer technical training in Python, and an immersive, project-based learning experience, focused on environmental and geospatial applications for faculty and undergraduates from partner institutions. Through participation in the EDSC summer internship, students and faculty from groups who have been historically underrepresented in STEM (e.g. women, people of color) gain hands-on experience with technical data science skills, and work in groups with peers on an applied EDS project, while developing collaboration and science communication skills, and growing their comfort and

confidence working with different data types. A core goal of the EDSC was to build institutional capacity to teach and learn EDS by offering technical and pedagogical professional development to faculty partners. The program also provided an opportunity for EDSC alumni to return to the summer program as advanced interns where they helped train their peers, and developed critical mentorship and leadership skills that supported at least one student to continue on to teach Python.

Conceptualized and funded prior to the onset of the COVID-19 pandemic, the EDSC was originally designed as a hybrid in-person and online program to begin in May 2020, where participants would spend a week on the University of Colorado campus in Boulder for in-person technical training, before returning to their home institutions for synchronous online instruction and project work time. However, in March 2020 the EDSC program was forced to abruptly move to a fully online model before in-person meetings were able to take place. Even though the virtual format allowed students to participate in the program while travel was restricted, and pandemic-related safety precautions were in place (Fletcher et al., 2021; Bawadi et al., 2023), online learners often report feeling isolated and detached from a larger, social classroom environment (Gillett-Swan, 2017) and struggle in navigating the technological demands (Bawadi et al., 2023). These challenges were addressed through clear scaffolding and ongoing guidance on interacting in virtual platforms, and through intentional group work and frequent check-ins.

Modeled after the Earth Analytics Data Science Bootcamp course (Palomino & Wasser, 2021), and informed by years of teaching EDS to professional graduate certificate students in a hybrid online and in-person setting, the EDSC fills a data

science training need by offering EDS education to faculty and undergraduates from historically underrepresented groups in a fully online learning environment, thereby accommodating participation without a need to relocate. Often students at MSIs are non-traditional, supporting families, jobs, and other responsibilities. While learning and teaching technical data science topics and skills online have their own unique challenges and constraints (Gulatee & Combes, 2008; Martin et al., 2021), we studied the impact of our approach to assess its effectiveness as a model. Similarly, we looked for evidence of growth in learners' science identity and sense of belonging in science, two crucial aspects of recruiting and retaining learners from marginalized communities in STEM (Dortch & Patel, 2017; Rainey et al., 2018; Rodriguez & Blaney, 2021). Below we describe our approach to EDS education and to democratizing access to data science through the EDSC. We share findings from data collected over three years of the EDSC and discuss how these results fit into the larger conversation around online learning environments, and teaching technical topics to learners from under-resourced communities. This work is framed by the following set of questions:

- *How does involvement in an immersive, online, project-based data science learning environment contribute to the development of participants':*
    - *Self-confidence in their Earth Data Science technical skills?*
    - *Science and data science communication skills?*
    - *Science identity and sense of belonging?*

Earth and Environmental Data Science Education

Earth and Environmental Data Science (EDS)

Our model of Earth and Environmental Data Science (EDS) refers to in-demand skills at the intersection of science and data science (Wasser et al, 2022). While data science and domain science have traditionally been taught separately, in today's data-driven world, there is a demand for those with both technical data science skills as well as domain science content knowledge (Pennington et al., 2020). Our instructional approach strengthens several components of EDS. First, we teach technical data science skills including scientific programming, version control, and the command line, which are critical skills for Earth data scientists (Hampton et al, 2017; Wasser et al, 2022). Next, domain science knowledge is needed as it supports data processing decisions, developing approaches to address scientific questions using data, and the selection of appropriate data needed to pose and address a research question. Understanding different data types is critical to our EDS education model, given that scientists spend nearly 80% of their time cleaning their data and preparing it for analysis (Snyder, 2019). Likewise, having a solid grasp of different data structures makes it easier and faster for learners to identify potential data issues, select appropriate processing tools and integrate data. Finally, communication and collaboration skills are essential for anyone working in EDS, given science is becoming increasingly transdisciplinary, and professional work environments are shifting towards remote options where collaboration often happens asynchronously.

Democratizing Access to EDS Education

There is a growing opportunity for learners to pursue careers in the booming data science job market (Manyika et al., 2017), yet learning technical data science skills

remains a challenge, since courses in these areas are traditionally offered at large, well-funded research-intensive universities (Tang & Sae-Lim, 2016). While successful models for teaching data science and computational methods in the health sciences (Tan, Elkin & Satagopan, 2022), applied statistics (Nolan & Temple Lang, 2015), and physics and astronomy (Caballero et al, 2019) have been reported, including summer workshops and research experience for undergraduates and faculty, there is a need for targeted training opportunities in the earth (Pennington et al, 2020) and environmental sciences (Emery et al, 2021). Our EDS education model is built on four core components that contribute to program scalability and sustainability: i.) provide EDS training to both undergraduate students and faculty at partner institutions, ii.) provide faculty with pedagogical support needed to integrate EDS curriculum into existing courses at their home institutions, iii.) empower the next generation of teachers by providing interns with the opportunity to return as advanced interns to mentor their peers and, iv.) widely democratize access to EDS curriculum by publishing teaching and learning materials online as Open Educational Resources (OER) that others including faculty at MSIs can use in their classrooms.

Throughout their participation in the EDSC summer program, faculty mentors participated in workshop sessions alongside undergraduate students and were a familiar point of contact for troubleshooting and support. As a faculty mentor, attending the workshops was a way to show students that faculty were involved and invested in their students' learning and success. Some faculty came in with EDS skills and others did not; however, all faculty grew from this experience and learned skills, techniques, and mentoring practices to incorporate into their classrooms and research. Faculty

mentors met together, provided feedback, offered troubleshooting support, shared resources, and worked towards broadening the reach of the EDSC through course material development. During their journey of acquiring experience in EDS, participating faculty dove into the field's complexity and its underlying components. In addition to grasping the effectiveness of data science, they ventured into the realm of advanced knowledge acquisition, which will soon necessitate a further exploration of cloud-based cyberinfrastructure (CI). Gaining a deeper understanding of CI and its interplay with EDS represents a crucial and demanding learning curve, which also surfaces the growing issue of data sovereignty. A significant milestone emerged with the initiation of the EDSC program's support for participating faculty in creating an EDS module designed for integration into their preferred courses. This not only enabled faculty members to seamlessly incorporate this module into one or more of their classes, but also facilitated the extension of summer projects into the academic year.

In years 2 and 3 of the program, we invited EDSC participants from the previous year to serve as advanced intern student mentors, creating and maintaining a channel of connectedness within the student group. It is vital for students to feel connected to peers who have succeeded, given we know vicarious experiences can be a highly valuable motivator for students, helping to build self-efficacy (Bandura, 1986; Ahern-Dodson et al., 2020; Lim et al., 2017; Fayram et al., 2018; Zaniewski & Reinholz, 2016). Following the model that we developed for our other EDS programs (Wasser et al, 2022), we published all learning resources for the EDSC online on our https://www.earthdatascience.org learning portal which sees over one million unique global users each year. The curriculum for the EDSC can be found in our Introduction to

Earth Data Science textbook (Palomino, Wasser & Joseph, 2021) and is available to be used for both asynchronous independent learning as well as embedded into existing courses.

The Earth Data Science Corps (EDSC)

Through a 12-week paid summer internship, the National Science Foundation (NSF)-funded EDSC included technical workshops and an immersive, project-based learning experience, intended to build participants' comfort and confidence with different technical and non-technical data science skills, provide experience with various data science practices, and help shape participants' science identities and sense of belonging to a larger community of data scientists. All EDSC workshops and meetings were virtual, taking place over Zoom (Zoom Video Communications, 2020) during the first year of the program (2020), and SpatialChat (SpatialChat, 2022) in subsequent years (2021, 2022). Homework assignments and other student questions were posted to a free and open-source Internet forum software called Discourse (Discourse, 2022), facilitating asynchronous communication between students and workshop instructors. Participants also made regular use of Slack (Slack, 2022), a messaging program designed for the professional workplace to maintain regular communication. While a JupyterHub (Jupyter, 2022) was maintained by Earth Lab staff during the first year of the EDSC to provide access to lesson notebooks, in the second and third years all programming and homework notebooks were completed using the free Google Colab environment (Google, 2022) which hosts cloud-based Jupyter Notebooks (Kluyver et al, 2016). Moving to Google Colab further builds capacity for faculty at MSIs to teach this material given the platform is also freely available. Faculty partners could then

seamlessly use the notebooks developed in our program in their classrooms in future years, allowing students to get started coding right away without the need to download packages or set up a local environment, a hurdle that can often be difficult to overcome, particularly for those who are new to programming (Kim & Henke, 2021). Google Colab also allowed participants to easily take their work with them at the end of the internship for continued access to course materials and lessons, without the need to install a working environment on their local computers.

Modeled after our Earth Data Analytics (EDA) Professional Graduate Certificate program at the University of Colorado Boulder (Wasser, Herwehe & Palomino, 2019) and the Data Carpentry teaching approach (Teal et al., 2015), EDSC technical workshops took place synchronously online 2-3 times each week for six weeks during the summer and covered topics including strings, lists, and operators in Python, working with tabular and time-series data using pandas (McKinney, 2011; Pandas Development Team, 2022), working with raster and vector spatial data using rioxarray (Snow et al, 2021; rioxarray contributors, 2023), GeoPandas (GeoPandas contributors, 2022) and EarthPy (Wasser et al, 2019), and plotting data using Matplotlib (Caswell et al., 2022). Participants were also introduced to topics related to writing clean, open, reproducible code that can be run on any operating system. EDSC workshop topics were briefly introduced by an EDSC instructor before participants worked with their peers and faculty mentors in small breakout sessions.

We strategically moved from Zoom to SpatialChat given our experiences teaching online in the EDA bootcamp at the start of the pandemic. Unlike Zoom, SpatialChat allows participants to move around a virtual online space. They move their

SpatialChat icons around freely in the room, to facilitate working together in small groups without hearing the online “chatter” of others in a room. SpatialChat further allows instructors to “drop in and visit” these small groups to answer questions as they would in a real classroom. This made for an interactive, online learning environment where users were encouraged to share their screens and engage in small groups and paired programming (Saltz & Heckman, 2020) to work through coding activities and homework challenges with their colleagues. The spatial component of SpatialChat also allowed participants to move around and zoom in or out to easily see everyone else in the room, and develop a better sense of how other participants were able to follow along through the use of emojis (e.g. thumbs up, confused faces, fire) to communicate with each other and the instructors.

After six weeks of technical workshops (Table 1), EDSC participants shifted their focus to a six-week, immersive, applied EDS group research project with peers from their home institutions and led by their faculty mentors (Table 2). During the six weeks of group project work time, teams met regularly (2-3 times/week) to organize project ideas and craft their project pitches using a Message Box (COMPASS Science Communication, 2023), to look for relevant data to answer their research questions, and develop workflows that applied Python and EDS skills learned during the technical training sessions. Throughout project work time, all of the teams would meet together as a large group to share weekly progress reports and solicit feedback from their peers. This gave each group the opportunity to practice their public speaking and science messaging in preparation for their final presentations that took place during the last week of the EDSC internship.

One critical aspect of successful project-based learning is a student's relationship to the topic being studied. Students and faculty worked together to select projects that were data-driven but also culturally relevant. Working in teams of 4-5 students and 1-2 faculty partners, projects from the EDSC focused on a variety of locally and culturally relevant topics that have included flooding on tribal lands, COVID-19 data sovereignty, developing spatial mapping tools for Tribal College campuses and communities, transportation, air-quality patterns, and water resource management in the Denver metropolitan area, bird behavior and methane fluxes in the Prairie Pothole Region in the Dakotas, among others. Student research projects included formal group presentations and written blog posts summarizing their methods and findings, giving participants an opportunity to further develop their science communication skills, a critical component of our EDS model (Wasser et al, 2022). Details about each EDSC group project can be found here: https://earthlab.colorado.edu/.../earth-data-science-corps-projects

Week 1: Software Carpentry	**Week 2:** Tool Introduction	**Week 3:** Python Fundamentals	**Week 4:** Tabular Data; Time-Series	**Week 5:** Vector & Raster Data	**Week 6:** Science Comms
Shell	Slack	Strings	Pandas dataframes	GeoPandas	Clean and reproducible code
Bash	Discourse	Lists	Datetime objects	EarthPy	Data visualization
Intro to Python	Google Drive	Operators	Parse dates	Rioxarray	Science collaboration
Version control with git & GitHub	Google Colab	Plotting with Matplotlib		CRS reprojection	
	Zoom/ SpatialChat			Crop; clip	

Table 1: Outline of tools and topics covered during the technical workshops (Weeks 1-6) of the Earth Data Science Corps summer internship.

Week 7: Project pitches	**Week 8:** Data presentations	**Week 9:** Methods presentations	**Week 10:** Initial results presentation	**Week 11:** Practice presentations	**Week 12:** Final presentations
Message box Initial project brainstorming Research question generating Groups look for relevant data Project pitch in short 5-10 min presentations	Groups share data in short 5-10 min presentations Complete checklist: *Where is the data from? When was it collected? How is data collected? What kind of data are you working with? How was the data accessed? How big is the data?*	Groups share methods used to address research question in short 5-10 min presentations Peer review Work on written blogs	Groups share initial results in short 5-10 min presentations Peer review Work on written blogs	Groups deliver practice presentations; 15-20 min Peer review Finalize written blogs	Groups deliver final presentations; 15-20 min Audience feedback Submit written blog

Table 2: Outline of topics covered during group project work time (Weeks 7-12) of the Earth Data Science Corps summer internship (2020-2022).

EDSC Demographics

The EDSC partnered with the University of Colorado Boulder, along with two Tribal Colleges and Universities (TCUs) and a Hispanic Serving Institution (HSI) for all three years of the program (2020-2022), as well as a local community college in the first year of the program (2020). These five institutions were selected for partnership prior to project funding to span a wide range of EDS education capacities and build a diverse, workforce-ready cohort of undergraduate students through training and applied projects. Faculty mentors recruited students from participating institutions and were encouraged to select applicants who could help build ethnic and gender diversity in the data sciences (e.g. women, minorities). Roughly one-third of participants reported having a deep connection with their Native tribal communities, with a nearly equal fraction collectively identifying as either African American, Asian, or Hispanic. Nearly two-thirds of students who participated in the EDSC identified as female (Figure 1). From 2020

through 2022, the EDSC trained three separate cohorts of participants, which included 61 undergraduate student interns and eight faculty mentors. In total, 53 participating undergraduate student interns provided consent to participate in this study (Year 1 [2020] = 19 students; Year 2 [2021] = 15 students; Year 3 [2022] = 19 students).

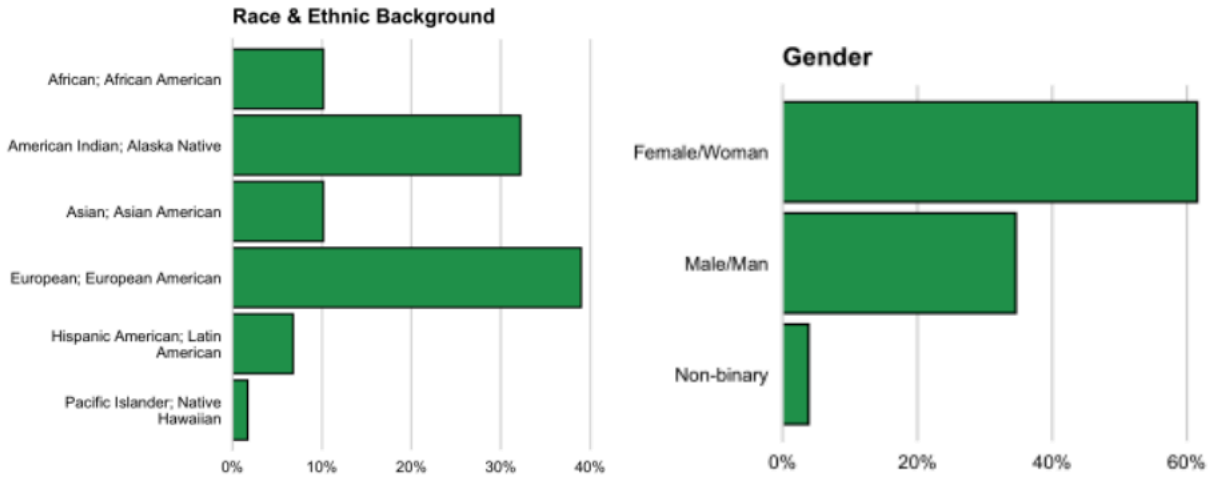

Figure 1: Demographic information from EDSC undergraduate student interns (2020-2022; *n* = 53). In six cases participants reported multiracial backgrounds.

Methods

Data Collection & Analysis

This study includes three years of survey data collected from participating EDSC undergraduate student interns to examine how the EDSC approach to data science instruction shaped learners' Python and data science skills, communication skills related to data science, confidence with different data science practices, and sense of identity and belonging in science. Likert surveys were developed (see Appendix Tables A1-A5) to assess changes across five different EDS constructs: technical i.) Python and ii.) data science skills, iii.) communication skills in science and data science, iv.) general data science practices, and v.) sense of science identity and belonging. The instruments intended to measure students' comfort and confidence with Python (Table A1) and data

science (Table A2) were designed by the EDSC assessment and evaluation team and included items that asked about creating reproducible workflows, publishing code, using Jupyter Notebooks, and working with time-series, tabular, spatial, and remotely sensed data using Python. Similarly, the authors developed survey items to assess EDSC participants' self-perceived confidence in communicating their science findings orally and through writing, including blogs (Table A3), and with different data science practices (Table A4) including generating research questions, and finding data to address those questions. In addition to measuring students' comfort and confidence with different technical data science and communication skills, this study also aimed to understand the effects on student identity and sense of belonging in science (Table A5). Questions asked as part of this dimension of the end-of-program survey were taken from a validated instrument used to measure science identity and sense of belonging in science (Chemers et al., 2011; Robnett et al., 2015).

Data were collected at the end of each of three summers (2020-2022) through a retrospective post-then-pre survey (Lam & Bengo, 2003) where participants provided two responses to each set of questions, one that measured their level of comfort or confidence *after* completing the EDSC workshops, and another that measured their comfort or confidence *before* participating in the EDSC. Of the different types of retrospective post-surveys, the retrospective post-then-pre design (sometimes written as retrospective pretest or post+retrospective pretest) has been shown to provide a more accurate measure of outcomes than the traditional pre-then-post design, and can lead to more accurate estimates of changes due to some intervention than a traditional pre-post design because it can control for response shift bias and result in a more

accurate participant reflection (Pratt et al., 2000; Drennan & Hyde, 2008; Cartwright & Atwood, 2014).

Analysis of Likert surveys traditionally involves assigning numerical values to each of the different response categories (e.g. 1 = *Strongly Disagree*, 2 = *Disagree*, 3 = *Neither Agree nor Disagree*, 4 = *Agree*, 5 = *Strongly Agree*). These numerical values are often totaled for each respondent and then used to calculate descriptive statistics (mean, standard deviation) and as variables in parametric statistical analysis (*t*-tests, ANOVA, etc.). This traditional approach to analyzing responses from Likert surveys, known as Classical Testing Theory (CTT), which makes use of raw responses, often violates normality requirements needed to carry out inferential statistics (Boone et al., 2013). Raw data collected from Likert surveys is ordinal in nature, and nothing is known about the distance between response categories like *Agree* and *Strongly Agree* (Boone, 2020). One technique used to convert ordinal data from Likert surveys to linear data that can be used in *t*-tests and analysis of variance (ANOVA) is Rasch modeling (Rasch, 1960). Part of a larger family of measurement approaches known as Item Response Theory (IRT), Rasch modeling is commonly used in psychometrics to assess the quality of survey instruments. This approach to measurement provides estimates of person ability, and item difficulty along a common log-odds (logit) scale. Here, a person's 'ability' can be thought of as the amount of an underlying trait held by that individual (e.g. level of Python or Data Science skill, etc.). Item difficulty is the level of a specific trait that a survey taker must have in order to endorse that particular survey item. A key feature of any Rasch model is the likelihood of a survey taker responding in a specific

way to a particular item is a function of the difference between the individual's ability and the difficulty of the survey item (Boone et al., 2013).

Ability and difficulty values measured in logits can also be reported as probabilities using the inverse logit function $e^x/(1 + e^x)$ where e is Euler's number (2.718) and x is the logit value (Abbakumov et al, 2020). When item difficulty and person ability are measured in percent probability, it provides a direct estimate of the likelihood of success (Wright & Stone, 1979). Higher values indicate a higher chance of success, while lower values indicate a lower chance of success. Person ability measured in percent probability represents the estimated probability that a particular individual will endorse an item of average difficulty. Item difficulty measured in percent probability represents how likely a respondent with average ability is to endorse a particular item.

For this study, a Rasch rating scale model (RSM; Andrich, 1978) for each of the five Likert-like survey instruments was built using the Test Analysis Modules package (TAM; Robitzsch et al., 2018) in the R programming environment (R Core Team, 2021). This package calculates measures of person ability, item difficulty, person and item separation reliability, as well as parameters used to identify misfitting items and assess unidimensionality (i.e. is the survey measuring a single construct?). Values of person ability obtained from the Rasch model were then used as variables in ANOVA and post-hoc *t*-testing to assess growth in participants' self-perceived comfort and confidence with different technical and non-technical data science skills following their participation in the EDSC program.

Results

Raw Response Percents (Before/After)

Plots of raw responses (as percentages) from each of the five surveys were developed to visually compare students' perceived comfort and confidence prior to and following their participation in the EDSC program (Figures 2-6). Based on these plots, students' comfort with specific Python skills (Figure 2), and confidence with general data science skills (Figure 3) underwent the greatest shift following instruction, with over 90% of respondents reporting either a complete lack of or only slight levels of comfort or confidence in some of these areas before instruction. Post-EDSC responses showed in most instances 60-80% of participants reported moderate to high levels of confidence or comfort with these same sets of technical Python and data science skills. While a shift in students' attitudes related to data science communication (Figure 4), data science practices (Figure 5), and science identity (Figure 6) were also observed in the plots of raw percentages, growth was less pronounced in these areas when compared to the shifts in responses reported on the surveys used to measure Python and data science skills. For each of these three survey instruments, at most 70-75% of respondents disagreed with or reported low levels of confidence on items asking about their data science communication, data science practices, or science identity and sense of belonging on the pre-instruction questions, suggesting that participants came into the program with a higher self-reported level of ability or comfort in these areas.

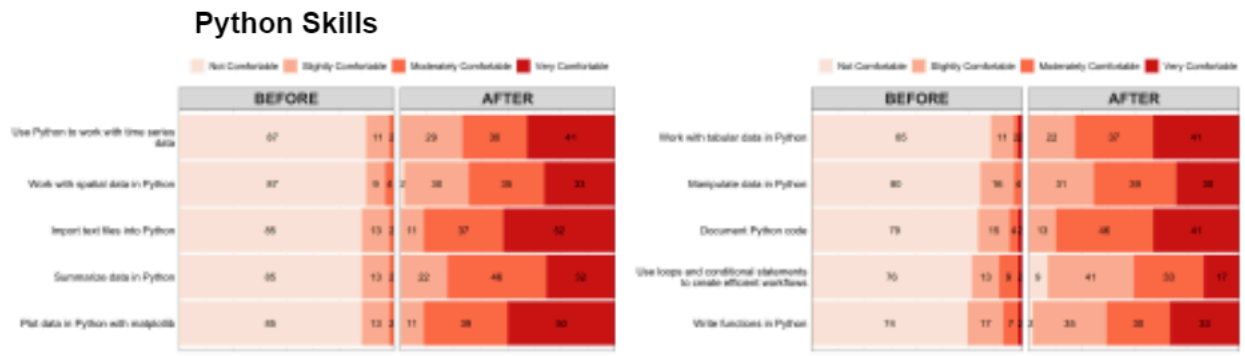

Figure 2: Distribution of raw responses (n = 46) to the Python Skills assessment items. Numerical labels indicate percent of respondents in each category.

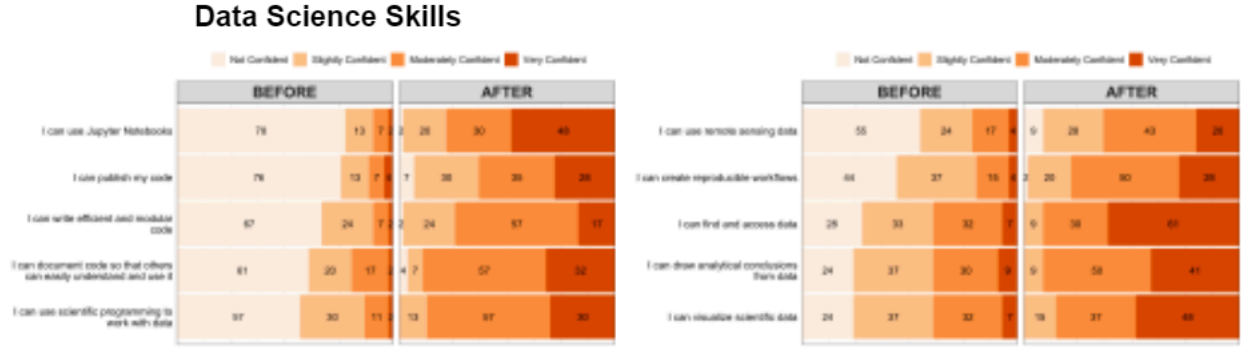

Figure 3: Distribution of raw responses (n = 46) to the Data Science Skills assessment items. Numerical labels indicate percent of respondents in each category.

Figure 4: Distribution of raw responses (n = 48) to the Data Science Communication Skills assessment items. Numerical labels indicate percent of respondents in each category.

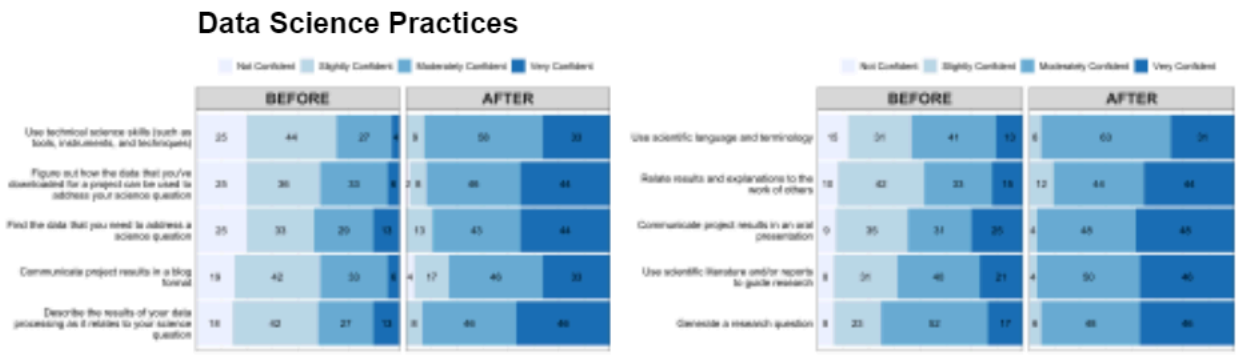

Figure 5: Distribution of raw responses (n = 48) to the Science Practices assessment items. Numerical labels indicate percent of respondents in each category.

Figure 6: Distribution of raw responses (n = 46) to the Science Identity assessment items (Robnett et al., 2015). Numerical labels indicate percent of respondents in each category.

Rasch Measurement

Raw survey responses were assigned numerical values based on category with 1 corresponding to the lowest level of endorsement (eg. *Not Confident*; *Not Comfortable*; *Strongly Disagree*) and a 4 or 5 denoting the highest level of endorsement (e.g. *Very Confident*; *Very Comfortable*; *Strongly Agree*) depending on the number of response categories. These values were used to develop rating scale Rasch models for each trial (before/after) of the five Likert survey instruments, and include measurements of person ability, item difficulty, item fit, and item and person separation reliability. Values of person ability ranged from -6.21 to 9.95 logits (0.20 to 99.9% probability) depending on trial (before/after) and survey dimension (Table 3). Measures of item difficulty were calculated for pre-instruction responses and then held constant as a means of

anchoring to post-instruction responses and ranged from -3.72 to 0.93 logits (2.36 to 71.7% probability) based on survey dimension (Table 4). Average item fit parameters (Boone, 2020) were calculated and found to be in an acceptable range of 0.7 to 1.3 (Bond et al., 2020) for both trials of the five survey dimensions (Table 5) with the exception of the post-instruction Python skills which over-fit the model. Reliability measures including person- and item-separation reliability were acceptable (> 0.50; Fisher, 2007) for all five survey dimensions, pre- and post-instruction (Table 5) indicating that both items and persons can be reliably measured using these different assessment tools.

Dimension	**Ability (logits; min-max; Before)**	**Ability (logits; min-max; After)**	**Ability (% probability; min-max; Before)**	**Ability (% probability; min-max; After)**
Python Skills	-0.20 – 6.56	3.92 – 9.95	45.0 – 99.8	98.1 – 99.9
Data Science Skills	-2.90 – 3.41	0.39 – 6.66	5.21 – 96.8	59.6 – 99.9
Data Science Communication	-2.75 – 2.47	0.46 – 5.79	6.01 – 92.2	61.3 – 99.7
Data Science Practices	-3.77 – 3.01	-0.68 – 5.24	2.25 – 95.3	33.6 – 99.5
Science Identity	-6.21 – 3.96	-3.36 – 3.96	0.20 – 98.1	3.36 – 98.1

Table 3: Min and max values of person ability measured in logits and % probability for both trials for each of the 5 survey dimensions.

Dimension	Item Difficulty (logits; min-max)	Item Difficulty (% probability; min-max
Python Skills	-0.98 – 0.93	27.3 – 71.7
Data Science Skills	-2.63 – -0.35	6.72 – 41.4
Data Science Communication	-3.01 – -0.01	4.69 – 49.8
Data Science Practices	-3.18 – -1.66	3.99 – 16.0
Science Identity	-3.72 – -2.56	2.36 – 7.18

Table 4: Min and max values of item difficulty measured in logits and % probability for each of the 5 survey dimensions.

Dimension	Trial	Average Item Infit (MNSQ)
Python Skills	*Before*	0.78
	After	1.83
Data Science Skills	*Before*	1.06
	After	0.99
Data Science Communication	*Before*	1.00
	After	1.29
Data Science Practices	*Before*	1.00
	After	0.79
Science Identity	*Before*	0.95
	After	1.27

Table 5: Measures of average mean-square (MNSQ) item infit for each trial of the 5 survey dimensions.

Dimension	Trial	Item (EAP) Reliability	Person (WLE) Reliability
Python Skills	*Before*	0.73	0.51
	After	0.86	0.83
Data Science Skills	*Before*	0.85	0.78
	After	0.84	0.81
Data Science Communication	*Before*	0.84	0.84
	After	0.80	0.76
Data Science Practices	*Before*	0.88	0.88
	After	0.80	0.76
Science Identity	*Before*	0.92	0.89
	After	0.84	0.77

Table 6: Measures of item (EAP) and person (WLE) reliability for the five dimensions assessed.

ANOVA / *t*-testing / effect size

To examine the role that cohort (year of the program), trial (before/after) and survey dimension played in the variation of measured values of person ability, ANOVA was carried out using person ability as the dependent variable, with trial, survey dimension, and cohort added to the model as independent variables (i.e. Person Ability ~ Trial + Dimension + Cohort). Results from ANOVA including effect sizes (partial eta squared [η_p^2]; Sink & Stroh, 2006) are summarized in Table 7. No statistically significant differences in person ability were present when comparing students from different cohorts of the EDSC ($F(2, 469) = 0.22$, $p = 0.81$, $\eta_p^2 = 0.0009$). Both trial ($F(1, 469) = 324.75$, $p < 0.001$, $\eta_p^2 = 0.41$) and survey dimension ($F(4, 469) = 45.56$, $p < 0.001$, $\eta_p^2 = 0.28$) contributed significantly to variation in values of person ability, with large effects ($\eta_p^2 > 0.14$) reported for each of these two factors. Post-hoc *t*-testing revealed significant growth in person ability from pre- to post-participation in the EDSC internship program across each of the five different survey dimensions (Table 8) when participants from all three cohorts were taken on the aggregate. Large effect sizes (Cohen's $d > 0.8$) were reported across four of the five survey dimensions, with science identity having a moderate effect ($t = 3.32$, $d = 0.69$). Looking at growth in participants' level of endorsement for each of the five survey dimensions across all three years of the program, significant gains in person ability were observed in each case with the exception of science identity in years 1 and 3 (Table 9). These results are reported graphically in Figure 7 which displays growth for all EDSC participants' ability across each of the five survey dimensions, and in Figure 8 which displays similar results faceted by cohort. Here, again, we see that of the five different survey dimensions, students' self-perceived Python, and data science skills underwent the largest growth

from pre- to post-participation in the EDSC across all three years of the program. Similar figures with values displayed on a percent probability scale have been included in the Appendix (Figure A1, Figure A2).

	Df	Sum Sq	Mean Sq	*F* value	*p* value	*Eta sq. (part)*
Trial	1	1004.44	1004.44	324.75	< 2e-16 ***	0.41
Dimension	4	563.71	140.93	45.56	< 2e-16 ***	0.28
Cohort	2	1.34	0.67	0.22	0.81	0.0009
Residuals	469	1422.77	3.09			

Table 7: Analysis of variance (ANOVA) table demonstrating how trial (before/after), survey dimension, and cohort (year of program) contribute to variation in measures of person ability. [* : $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$]

Dimension	*t*	Cohen's *d*
Python Skills	15.43 ***	3.22
Data Science Skills	10.45 ***	2.18
Data Science Communication	10.27 ***	2.09
Data Science Practices	7.50 ***	1.53
Science Identity	3.32 **	0.69

Table 8: Results from post-hoc *t*-testing and corresponding effect size values aggregated across all EDSC student participants (Cohen's *d*). [* : $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$]

Dimension	Cohort	*t*	*n*
Python Skills	2020	10.95 ***	16
	2021	7.69 ***	12
	2022	8.84 ***	18
Data Science Skills	2020	5.68 ***	16
	2021	5.68 ***	12
	2022	6.56 ***	18
Data Science Communication	2020	5.48 ***	16
	2021	6.39 ***	13
	2022	6.30 ***	19
Data Science Practices	2020	3.42 **	16
	2021	5.89 ***	13
	2022	4.13 ***	19
Science Identity	2020	1.60	16
	2021	2.42 *	12
	2022	1.83	18

Table 9: Results from post-hoc *t*-testing for each of the five survey dimensions across each of the three cohorts of EDSC students. [* : $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$]

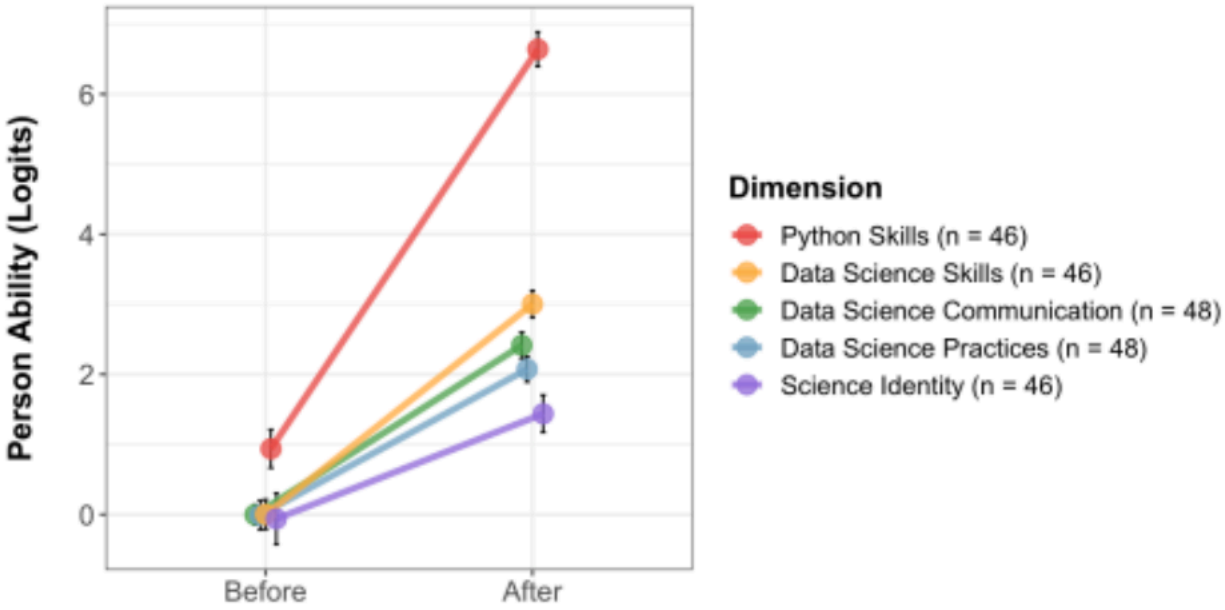

Figure 7: Growth in measures of person ability (logits) for the five different dimensions across the two timepoints aggregated across all EDSC student participants (before and after participation in the EDSC program).

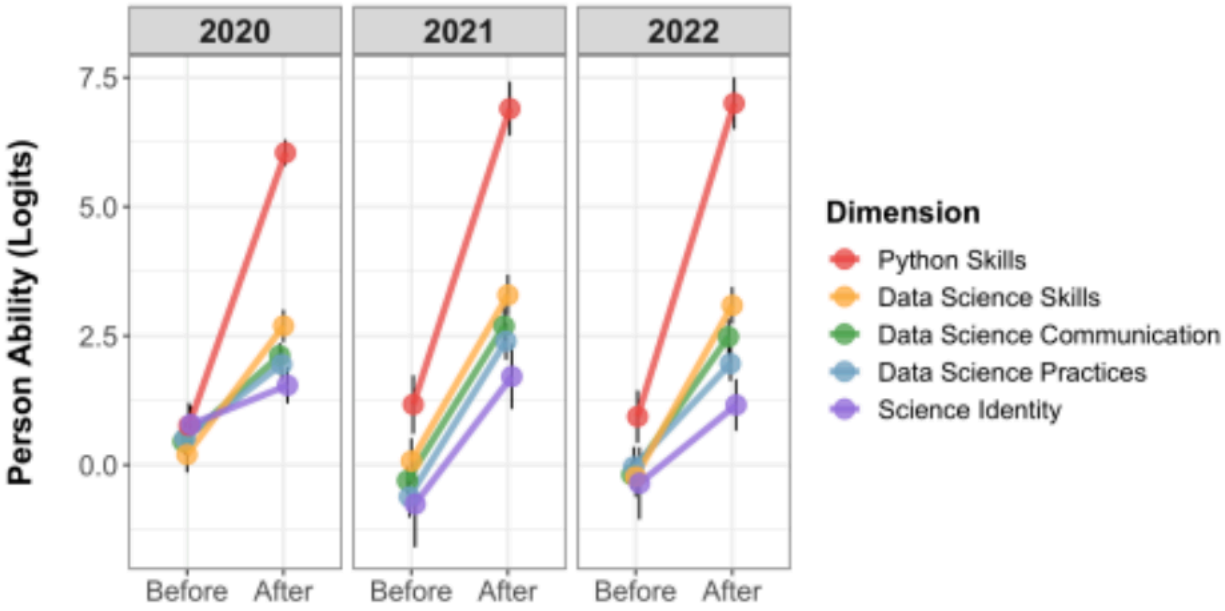

Figure 8: Growth in measures of person ability (logits) for the five different dimensions across the two timepoints for each of the 3 cohorts of EDSC students (before and after participation in the EDSC program).

Discussion

The core goal of the NSF-funded EDSC program was to provide training for students and faculty from communities that have been historically underrepresented in STEM through hands-on, project-based EDS training. Partnering with TCUs and HSIs, we helped participants develop comfort with different areas of EDS and self-identify as members of the EDS community. Over the course of 12 weeks in the summertime from 2020-2022, three cohorts of undergraduate interns (60 total) and faculty mentors (8 total) participated in virtual technical EDS training using Python, before going on to apply those skills to culturally relevant projects of their own design. This effort is situated within the larger objectives of the Harnessing the Data Revolution (HDR) Data Science Corps program, from one of the NSF's 10 Big Ideas which aims to provide practical experiences, teach new skills, and offer learning opportunities, in a variety of settings, to data scientists and data science students.

We demonstrated significant shifts in EDS skills and self-identification as EDS scientists among the group of EDSC undergraduate interns, and we argue that both are critical for advancing learning and teaching data skills for students from underrepresented populations in STEM. Significant growth in participants' self-reported comfort and confidence with different components of their data science toolkits included data science communication and data science practices, as well as their Python and data science skills as measured across all three cohorts of EDSC student participants (Figure 8, Table 9). Through the EDSC program, we demonstrated that our approach to online data science education resulted in substantial gains in self-reported comfort and confidence in different technical data science areas in a group of novice programmers from historically marginalized communities.

Despite limitations associated with the sudden shift to emergency online learning during the COVID-19 pandemic, (Aguilera-Hermida, 2020), we found that when an emphasis was placed on communication, accessibility, and community building, significant growth in participants' self-reported technical and non-technical data science skills, data science practices, and fostering of science identities was achieved in the online learning space. Our data shows that having consistent communication through multiple channels including Slack, Discourse, Zoom, and SpatialChat, making sure teaching and learning materials are open and accessible through our https://www.earthdatascience.org learning portal, and working towards building a community of learners are all critically important when working with beginner programmers, particularly those who may also lack internet bandwidth or consistent internet connectivity.

Overall, our results show that an approach to online data science education that relies on open and accessible educational resources, emphasizes collaborative team work environments, and empowers participants to take control of their own learning through team-designed applied projects, has the ability to significantly shift learners' comfort and confidence with technical and non-technical data science skills. Importantly, our approach to online teaching is adaptable to other domain-specific fields of data science, as well as data science education more broadly, and transferable to other disciplines beyond earth and environmental data science.

Program sustainability and scalability

The following aspects are key to program sustainability and scalability: i.) capacity building to teach and learn EDS at participating institutions, ii.) supporting EDSC alumni through the advanced internship program, iii.) a commitment to open science through the use of tools like Google Colab and publishing EDSC curricular materials on our free and open EDS learning platform (https://www.earthdatascience.org/), and iv.) emerging partnerships with nationally serving organizations focused on undergraduate education and STEM faculty development. Through professional development, participating EDSC faculty mentors developed EDS modules and curricular materials that they could use at their home institutions, to begin incorporating EDS, Python, and Jupyter Notebooks into their existing GIS, geography, and ecology courses. As faculty partners carry content and pedagogy learned through EDSC participation to their classrooms, they will continue teaching EDS skills to their students, sustaining and scaling the impact of this program.

The advanced internship component of the EDSC allowed alumni to return in years 2 and 3 of the program to develop peer mentoring experience, continue adding to their data science toolkits, and build their capacity to teach EDS content to others. Several EDSC alumni who took part in the advanced intern program have presented their work at professional conferences, sought out additional internship opportunities, matriculated into graduate degree programs focused on data analytics, and returned to their home institutions to teach data science courses. These and other EDSC alumni contribute to the foundation of the next generation of diverse earth and environmental data science practitioners, and they will help sustain the message and vision of the EDSC for many years to come.

We see our commitment to open, reproducible science through the use of tools like Google Colab and our online EDS learning platform (https://www.earthdatascience.org/) as a primary sustainability engine for the EDSC program. Free, cloud-based development platforms like Google Colab allow participants to get started coding right away without having to overcome challenging hurdles like installing Python environments on their local computers, and provide continued access to learning materials after the conclusion of the internship. Similarly, the resources contained within our online learning portal are free, licensed through Creative Commons Attribution ShareALike (https://www.earthdatascience.org/license/), accessed by millions of users each year, and are available for teachers and learners to modify and meet their own unique needs. Through our online learning portal, EDSC interns and faculty partners have unlimited access to the EDSC curriculum after completion of the internship, and are encouraged to use these resources in their classrooms and as

reference materials as they continue adding to their EDS skill sets, and incorporating EDS modules into their teaching and research.

Lastly, partnerships that have emerged as a result of the EDSC with nationally recognized organizations focused on undergraduate education and faculty development are another mechanism that helps sustain and scale similar programs that are inspired by and modeled after the EDSC internship. One noteworthy example has been the formation of a STEM faculty working group at TCUs, spearheaded by collaboration with the American Indian Higher Education Consortium (AIHEC). The primary objectives of this working group include introducing EDS concepts to Indian Country by training faculty at TCUs to incorporate EDS in their STEM curriculum including the use of LiDAR and other remotely sensed data, facilitating partnerships with TCU faculty beyond those who participated in the EDSC, and empowering TCUs and local tribes to develop and maintain cloud-based data management workflows. EDSC also led to a partnership with the NSF-funded Macrosystems Ecology For All Research Coordination Network (MEFA), a spin-off of the Ecological Research as Education Network (EREN; Stack Whitney et al, 2022; Styers et al, 2021). Through this emerging partnership, we are creating a series of training opportunities, modeled after the faculty development component of the EDSC, specifically designed for undergraduate biology and environmental science faculty interested in incorporating EDS modules into their existing curriculum.

Efforts like those described above to build capacity, train the next generation of EDS educators, make curricular materials freely available, and develop partnerships with nationally serving organizations focused on faculty development help with program

scalability and sustainability. However, internship programs that compensate students and instructors are reliant on funding sources. In order to fully harness the data revolution, and continue building diversity into the future of the EDS workforce, it is critical to continue offering such program opportunities. One example of these continued efforts is through a new national synthesis center, the Environmental Data Science Innovation and Inclusion Lab, ESIIL (Balch et al, 2022). This NSF-funded data synthesis center, hosted at the University of Colorado Boulder is built on four key pillars: i.) team science approaches and research in environmental data science, ii.) innovative tools and collaborative cyberinfrastructure, iii.) building inclusive partnerships and groups of earth and environmental data scientists, and iv.) data science education and training programs for a diverse and inclusive community. The ESIIL Stars program, modeled after the EDSC, is a five-month paid internship that partners with communities from underrepresented groups in STEM, and combines online data skills training, career development, an open textbook, and project based learning for undergraduates and faculty from MSIs.Through the ESIIL Stars program, we will continue to refine our approach to EDS education, and further develop partnerships that emerged out of the EDSC.

Future directions

While this study paints a positive picture of one approach to learning and teaching EDS online, improvements in future work include incorporating survey instruments that do not rely on self-reporting, expanding the participants to a larger, more randomized group, and valuing important anecdotal stories around learning successes. For example, finding a more objective measure of Python proficiency or

data science skills would improve our understanding. While we have attempted to control for participants' response shift bias through the use of a retrospective post-then-pre design (Marshall et al., 2007), there is potential for self-reporting response bias. Future work should develop objective measures to assess different components of learners' technical and non-technical data science skills. Further, there is also potential for a larger, randomly selected group of participants in order for these findings to be generalized to the larger population of data science learners. Also, anecdotal stories about the learning that took place are critical to capture. During the technical workshops, and particularly during group project work time, participants learned to grapple with large, messy, heterogeneous data sets, and deal with issues related to Indigenous Data Sovereignty. Participants also found innovative ways of making the data science tools and techniques that were developed as part of the EDSC work alongside the approaches to doing EDS that were already critical components of their data science toolkits, including using proprietary software and other programming languages not taught as part of the EDSC. These learning experiences are important for the EDSC story, and ones that we hope to continue to tell in research efforts going forward.

This study provides a successful model for the broader community of data science educators and researchers who are seeking ways to make their classrooms more accessible, looking for tools to assess data science learners' comfort and confidence, and working on developing assessment and evaluation tools to better understand data science learning in other contexts. Through this work, we have demonstrated that the skills being taught in the EDSC framework can be blended to

support relevant, culturally responsive, and meaningful project-based learning. What makes the EDSC unique is its purposeful approach to supporting smaller MSIs, including TCUs, to build capacity to learn and teach technical data science skills on their home campuses, a critical component of helping to create a more diverse STEM and data science workforce, and a mechanism for sustaining and scaling the teachings of the EDSC. While our research focuses on learning data science in the realm of solving applied EDS challenges, we see this approach as being applicable to STEM and data science educators on a broader scale. We developed this approach to be something that others can model and implement at their campuses and we welcome conversations with groups who are interested in partnering with us as we work to harness the data revolution and help build the next generation of earth and environmental data scientists.

Acknowledgements

This work was supported through NSF Awards #1924337 under the Harnessing the Data Revolution (HDR) Big Idea, and DBI-2153040 under the Environmental Data Science Innovation and Inclusion Lab (ESIIL). This research was supported in part by the NOAA cooperative agreements NA17OAR4320101 and NA22OAR4320151.The program was led by the Earth Analytics Education Initiative at Earth Lab, part of the Collaborative Institute for Research in the Environmental Sciences (CIRES) and supported by the Grand Challenge at University of Colorado Boulder.

Data Availability Statement

Data and code used to do analysis for this work have been deidentified and are currently stored in a public GitHub repository (https://github.com/earthlab/2022-edsc-manuscript).

IRB Statement

The surveys used to collect data for this manuscript were reviewed and approved by the IRB committee at the University of Colorado - Boulder (20-0254) in accordance with federal regulation and the Belmont Report principles. Consent was obtained from all study participants as required in our IRB-approved protocol.

References

Abbakumov, D., Desmet, P., & Van den Noortgate, W. (2020). Rasch model extensions for enhanced formative assessments in MOOCs. Applied Measurement in Education, 33(2), 113-123.

Aguilera-Hermida, A. P. (2020). College students' use and acceptance of emergency online learning due to COVID-19. International Journal of Educational Research Open, 1, 100011.

Ahern-Dodson, J., Clark, C. R., Mourad, T., & Reynolds, J. A. (2020). Beyond the numbers: understanding how a diversity mentoring program welcomes students into a scientific community. Ecosphere, 11(2), e03025.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43(4), 561-573.

Balch, J., Nagy, R., Amaral, C., Culler, E., Gold, A. U., Iglesias, V., ... & Quarderer, N. (2022, December). The Environmental Data Science Innovation & Inclusion Lab (ESIIL): a next-generation NSF data synthesis center. In AGU Fall Meeting Abstracts (Vol. 2022, pp. ED12C-0371).

Bandura, A. (1986). Social foundations of thought and action. Englewood Cliffs, NJ, 1986(23-28).

Bawadi, H., Shami, R., El-Awaisi, A., Al-Moslih, A., Abdul Rahim, H., Du, X., ... & Al-Jayyousi, G. F. (2023). Exploring the challenges of virtual internships during the COVID-19 pandemic and their potential influence on the professional identity

of health professions students: A view from Qatar University. Frontiers in medicine, 10, 1107693.

Bond, T.G., Yan, Z., & Heene, M. (2020). Applying the Rasch Model: Fundamental Measurement in the Human Sciences (4th ed.). Routledge. https://doi.org/10.4324/9780429030499

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). Rasch analysis in the human sciences. Springer Science & Business Media.

Boone, W. J. (2020). Rasch basics for the novice. In Rasch Measurement (pp. 9-30). Springer, Singapore.

Caballero, M. D., Chonacky, N., Engelhardt, L., Hilborn, R. C., del Puerto, M. L., & Roos, K. R. (2019). PICUP: A community of teachers integrating computation into undergraduate physics courses. The Physics Teacher, 57(6), 397-399.

Cartwright, T. J., & Atwood, J. (2014). Elementary pre-service teachers' response-shift bias: Self-efficacy and attitudes toward science. International Journal of Science Education, 36(14), 2421-2437.

Caswell, T. A., Lee, A., Droettboom, M., de Andrade, E. S., Hoffmann, T., Klymak, J., … Kniazev, N. (2022). matplotlib/matplotlib: REL: v3.6.0rc2 (Εκδοχή v3.6.0rc2). doi:10.5281/zenodo.7032953

Chemers, M.M., Zurbriggen, E.L., Syed, M., Goza, B.K. & Bearman, S. (2011). The Role of Efficacy and Identity in Science Career Commitment Among Underrepresented Minority Students. Journal of Social Issues, 67: 469-491. https://doi.org/10.1111/j.1540-4560.2011.01710.x

COMPASS Science Communication. (2023). https://www.compasssscicomm.org/leadership-development/the-message-box/. Accessed Sept. 18, 2023.

Discourse. (2022). Discourse Community Forum. https://www.discourse.org/

Dortch, D., & Patel, C. (2017). Black undergraduate women and their sense of belonging in STEM at predominantly White institutions. NASPA Journal About Women in Higher Education, 10(2), 202-215.

Drennan, J., & Hyde, A. (2008). Controlling response shift bias: the use of the retrospective pre-test design in the evaluation of a master's programme. Assessment & Evaluation in Higher Education, 33(6), 699-709.

Fayram, J., Boswood, N., Kan, Q., Motzo, A., & Proudfoot, A. (2018). Investigating the benefits of online peer mentoring for student confidence and motivation. International Journal of Mentoring and Coaching in Education, 7(4), 312-328.

Fisher, W. P. (2007). Rating Scale Instrument Quality Criteria. Rasch Measurement Transaction, 21 (1), 1095.

Fletcher Jr, E. C., Minar, N. J., & Rice, B. A. (2023). The Future Ready Lab: Maintaining Students' Access to Internships during Times of Crisis. Education and Urban Society, 55(5), 577-592.

GeoPandas contributors. (2023). geopandas/geopandas: v0.13.2 (v0.13.2). Zenodo. https://doi.org/10.5281/zenodo.8009629

Gibert, K., Horsburgh, J. S., Athanasiadis, I. N., & Holmes, G. (2018). Environmental data science. Environmental Modelling & Software, 106, 4-12.

Gillett-Swan, J. (2017). The challenges of online learning: Supporting and engaging the isolated learner. Journal of Learning Design, 10(1), 20-30.

Google. (2022). Welcome to Colaboratory. Retrieved September 6, 2022, from https://colab.research.google.com/

Gulatee, Y., & Combes, B. (2008). Identifying social barriers in teaching Computer Science topics in a wholly online environment. Science Mathematics and Technology Education: Beyond Cultural Boundaries, 173.

Hampton, S. E., Jones, M. B., Wasser, L. A., Schildhauer, M. P., Supp, S. R., Brun, J., ... & Aukema, J. E. (2017). Skills and knowledge for data-intensive environmental research. BioScience, 67(6), 546-557.

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., & Sethupathy, G. (2016). The age of analytics: Competing in a data-driven world. McKinsey Global Institute, 4, 136.

Iglesias, V., Braswell, A. E., Rossi, M. W., Joseph, M. B., McShane, C., Cattau, M., ... & Travis, W. R. (2021). Risky development: Increasing exposure to natural hazards in the United States. Earth's future, 9(7), e2020EF001795.

Jupyter. (2022). Retrieved September 7, 2022, from https://jupyter.org/hub

Kim, B., & Henke, G. (2021). Easy-to-use cloud computing for teaching data science. Journal of Statistics and Data Science Education, 29(sup1), S103-S111.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... & Willing, C. (2016). Jupyter Notebooks-a publishing format for reproducible computational workflows. Elpub, 2016, 87-90.

Lam, T. C., & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. American Journal of Evaluation, 24(1), 65-80.

Lim, J. H., MacLeod, B. P., Tkacik, P. T., & Dika, S. L. (2017). Peer mentoring in engineering:(un) shared experience of undergraduate peer mentors and mentees. Mentoring & Tutoring: Partnership in Learning, 25(4), 395-416.

Marshall, J. P., Higginbotham, B. J., Harris, V. W., & Lee, T. R. (2007). Assessing program outcomes: Rationale and benefits of posttest-then-retrospective-pretest designs. Journal of Youth Development, 2(1), 118-123.

Martin, F., Shanley, N. E., Hite, N., Pérez-Quiñones, M. A., Ahlgrim-Delzel, L., Pugalee, D., & Hart, E. (2021). High School Teachers Teaching Programming Online: Instructional Strategies Used and Challenges Faced.

Manyika, J., Chui, M., Madgavkar, A. & Lund, S. (2017). Technology, jobs and the future of work. San Fran

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 14(9), 1-9.

Monteleoni, C., Schmidt, G. A., & McQuade, S. (2013). Climate informatics: accelerating discovering in climate science with machine learning. Computing in Science & Engineering, 15(5), 32-40.

Nolan, D., & Temple Lang, D. (2015). Explorations in statistics research: An approach to expose undergraduates to authentic data analysis. The American Statistician, 69(4), 292-299.

Palomino, J., & Wasser, L. (2021). earthlab/earth-analytics-bootcamp-course: Bootcamp course 2.0 (2.0). Zenodo. https://doi.org/10.5281/zenodo.5418486

Palomino, J., Wasser, L., & Joseph, M. (2021). earthlab/earth-analytics-intro-to-earth-data-science-textbook: Earth Analytics Updated Version of the Intro Textbook (1.5). Zenodo. https://doi.org/10.5281/zenodo.4686073

Pandas Development Team. (2022). pandas-dev/pandas: Pandas (Εκδοχή v1.5.0rc0). doi:10.5281/zenodo.7018966

Pennington, D., Ebert-Uphoff, I., Freed, N., Martin, J., & Pierce, S. A. (2020). Bridging sustainability science, earth science, and data science through interdisciplinary education. Sustainability Science, 15(2), 647-661.

Pratt, C. C., McGuigan, W. M., & Katzev, A. R. (2000). Measuring program outcomes: Using retrospective pretest methodology. American Journal of Evaluation, 21(3), 341-349.

Quarderer, N., Halama, K., Post, A. K., Hummel du Amaral, C., Culler, E., Nagy, R. C., Tuff, T., Sanovia, J., Balch, J., Rattling Leaf, J., Gold, A., Swetnam, T. L., Monteleoni, C., Parker, J., Sullivan, S., & Iglesias, V. (2023). Creating Inclusive Spaces for Earth and Environmental Data Science Education (Version 2). figshare. https://doi.org/10.6084/m9.figshare.23990001.v2

R Core Team. (2021). R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Rainey, K., Dancy, M., Mickelson, R., Stearns, E., & Moller, S. (2018). Race and gender differences in how sense of belonging influences decisions to major in STEM. International Journal of STEM Education, 5(1), 1-14.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

rioxarray contributors. (2023). rioxarray: 0.15.0 Release (0.15.0). Zenodo. https://doi.org/10.5281/zenodo.8247542

Robitzsch, A., Kiefer, T., & Wu, M. (2018). TAM: Test analysis modules. R package version 3.0-21. https://CRAN.R-project.org/package=TAM.

Robnett, R. D., Chemers, M. M., & Zurbriggen, E. L. (2015). Longitudinal associations among undergraduates' research experience, self efficacy, and identity. Journal of Research in Science Teaching, 52(6), 847-867.

Rodriguez, S. L., & Blaney, J. M. (2021). "We're the unicorns in STEM": Understanding how academic and social experiences influence sense of belonging for Latina undergraduate students. Journal of Diversity in Higher Education, 14(3), 441.

Saltz, J., & Heckman, R. (2020). Using Structured Pair Activities in a Distributed Online Breakout Room. Online Learning, 24(1), 227-244.

Sink, C. A., & Stroh, H. R. (2006). Practical significance: The use of effect sizes in school counseling research. Professional School Counseling, 401-411.

Slack. (2022). Slack. https://www.slack.com

Snow, A. D., Brochart, D., Chegini, T., Hoese, D., Hamman, J., RichardScottOZ, … Pmallas. (2021). corteva/rioxarray: 0.3.1 Release (Εκδοχή 0.3.1). doi:10.5281/zenodo.4570457

Snyder, J. (2019). Data cleansing: an omission from data analytics coursework. Information Systems Education Journal, 17(6), 22.

SpatialChat. (2022). SpatialChat. [Software]. https://www.spatial.chat/

Stack Whitney, K., Heard, M. J., Anderson, L. J., Cooke, S., Garneau, D., Kilgore, J., ... & Parker, A. T. (2022). Flexible and Inclusive Ecology Projects that Harness Collaboration and NEON-Enabled Science to Enhance Student Learning. The Bulletin of the Ecological Society of America, 103(2), e01963.

Styers, D. M., Schafer, J. L., Kolozsvary, M. B., Brubaker, K. M., Scanga, S. E., Anderson, L. J., ... & Barnett, D. (2021). Developing a flexible learning activity on biodiversity and spatial scale concepts using open-access vegetation datasets from the National Ecological Observatory Network. Ecology and Evolution, 11(9), 3660-3671.

Tan, K. S., Elkin, E. B., & Satagopan, J. M. (2022). A model for an undergraduate research experience program in quantitative sciences. Journal of Statistics and Data Science Education, 30(1), 65-74.

Tang, R., & Sae-Lim, W. (2016). Data science programs in US higher education: An exploratory content analysis of program description, curriculum structure, and course focus. Education for Information, 32(3), 269-290.

Teal, T. K., Cranston, K. A., Lapp, H., White, E., Wilson, G., Ram, K., & Pawlik, A. (2015). IJDC| General Article. International Journal of Digital Curation, 10(1), 135-143.

Wasser, L., Palomino, J., Herwehe, L., Quarderer, N., McGlinchy, J., Balch, J., & Joseph, M. B. (2022, February 24). Student-Directed Learning in the Open Earth & Environmental Data Science Classroom. https://doi.org/10.31219/osf.io/xdj4z

Wasser, L. A., Herwehe, L., & Palomino, J. (2019, December). Democratizing Access to Earth Data Science Skills Using Blended Online and In-Person Approaches and Open Education. In AGU Fall Meeting Abstracts (Vol. 2019, pp. ED13D-0905).

Wasser, L., Joseph, M. B., McGlinchy, J., Palomino, J., Korinek, N., Holdgraf, C., & Head, T. (2019). EarthPy: A Python package that makes it easier to explore and plot raster and vector data using open source Python tools. Journal of Open Source Software, 4(43), 1886.

Wright, B. D., & Stone, M. H. (1979). Best test design.

Zaniewski, A. M., & Reinholz, D. (2016). Increasing STEM success: a near-peer mentoring program in the physical sciences. International Journal of STEM Education, 3(1), 1-12.

Zoom Video Communications. (2020). Zoom. [Software]. https://zoom.us

Appendix (Survey Instruments)

Please provide two responses for each statement. We want to understand your comfort doing the following tasks now AFTER you finished participating in the Earth Data Science Corps program compared to BEFORE participating in the program.	
Q1 - Import text files into Python	Q6 - Plot data in Python with matplotlib
Q2 - Summarize data in Python	Q7 - Document Python code
Q3 - Manipulate data in Python	Q8 - Use Python to work with time-series data
Q4 - Write functions in Python	Q9 - Work with tabular data in Python
Q5 - Use loops and conditional statements to create efficient workflows	Q10 - Work with spatial data in Python

Table A1: Items 1-10 of the **Python Skills** survey instrument. Responses were provided using a 4 point Likert scale of comfort from *Not Comfortable* to *Very Comfortable*.

Please provide two responses for each statement to tell us how confident you feel in doing these tasks AFTER you finished participating in the Earth Data Science Corps program compared to BEFORE participating in the program.	
Q1 - I can find and access data	Q6 - I can draw analytical conclusions from data
Q2 - I can create reproducible workflows	Q7 - I can visualize scientific data
Q3 - I can use scientific programming to work with scientific data	Q8 - I can document code so others can easily understand and use it
Q4 - I can write efficient and modular code	Q9 - I can publish my code
Q5 - I can use remote sensing data	Q10 - I can use Jupyter Notebooks

Table A2: Items 1-10 of the **Data Science Skills** survey instrument. Responses were provided using a 4 point Likert scale of confidence from *Not Confident* to *Very Confident*.

Please provide two responses for each statement to indicate to what extent you agree or disagree. We want to understand your comfort doing the following tasks now AFTER you finished participating in the Earth Data Science Corps program compared to BEFORE participating in the program.	
Q1 - I feel comfortable using data in spreadsheets	Q6 - I can create plots and maps to visual data using Python
Q2 - I am comfortable analyzing data	Q7 - When I write about science, I can present my ideas clearly
Q3 - I am comfortable writing code to analyze data	Q8 - I am confident I can communicate about science by writing blogs
Q4 - I know how to document formulas/code so others can understand and use it	Q9 - I am confident in my ability to explain science to non-scientists
Q5 - With enough time I feel I can learn new programming languages	Q10 - I am confident in my ability to collaborate with others on projects

Table A3: Items 1-10 of the **Data Science Communication** survey instrument. Responses were provided using a 5 point Likert scale of agreement from *Strongly Disagree* to *Strongly Agree*.

For the questions below, please provide two responses for each statement about your confidence in your abilities to work as a scientist and to complete the following tasks. We would like to understand your confidence AFTER you have completed the Earth Data Science Corps program, compared to your confidence BEFORE the program.	
Q1 - Use technical science skills (such as tools, instruments, and techniques)	Q6 - Describe the results of your data processing as it relates to your science question
Q2 - Use scientific language and terminology	Q7 - Use scientific literature and/or reports to guide research
Q3 - Generate a research question	Q8 - Relate results and explanations to the work of others
Q4 - Find the data that you need to address a science question	Q9 - Communicate project results in an oral presentation
Q5 - Figure out how the data that you've downloaded for a project can be used to address your science question	Q10 - Communicate project results in a blog format

Table A4: Items 1-10 of the **Data Science Practices** survey instrument. Responses were provided using a 4 point Likert scale of confidence from *Not Confident* to *Very Confident*.

Please provide two responses for each statement to indicate to what extent you agree or disagree. We want to understand how much you think being a scientist is part of who you are now AFTER you finished participating in the Earth Data Science Corps program compared to BEFORE participating in the program.	
Q1 - In general, being a scientist is an important part of my self-image	Q4 - I have come to think of myself as a "scientist"
Q2 - I have a strong sense of belonging to the community of scientists	Q5 - I feel like I belong in the field of science
Q3 - Being a scientist is an important reflection of who I am	

Table A5: Items 1-5 of the **Science Identity** survey instrument (adapted from Robnett et al., 2015). Responses were provided using a 5 point Likert scale of agreement from *Strongly Disagree* to *Strongly Agree*.

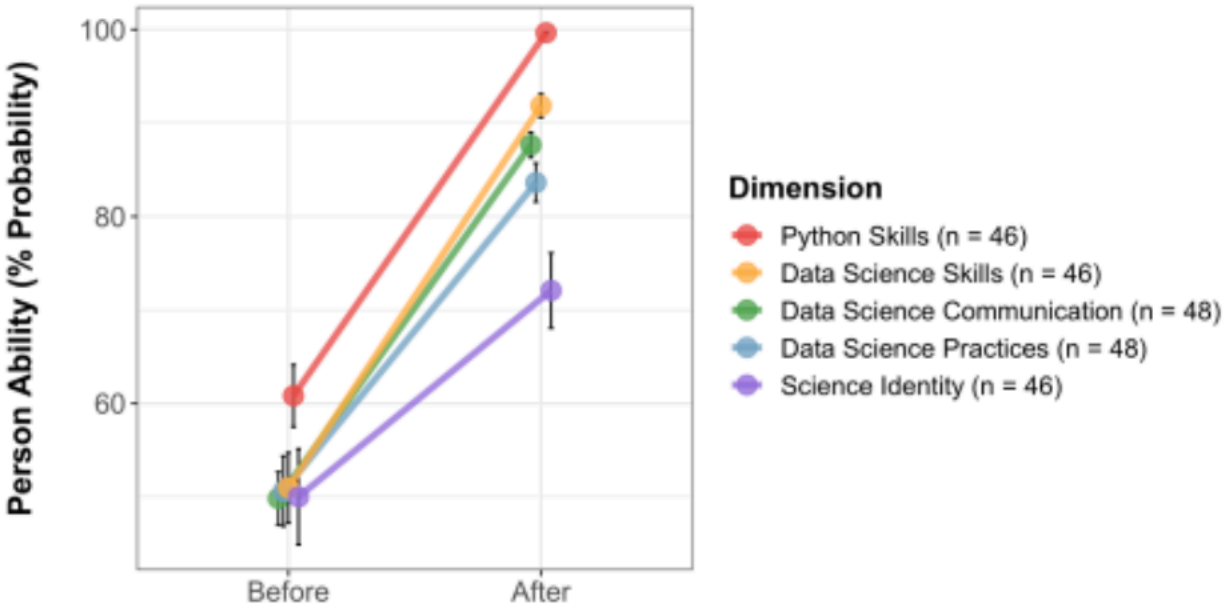

Figure A1: Growth in measures of person ability (% probability) for the five different dimensions across the two timepoints aggregated across all EDSC student participants (before and after participation in the EDSC program).

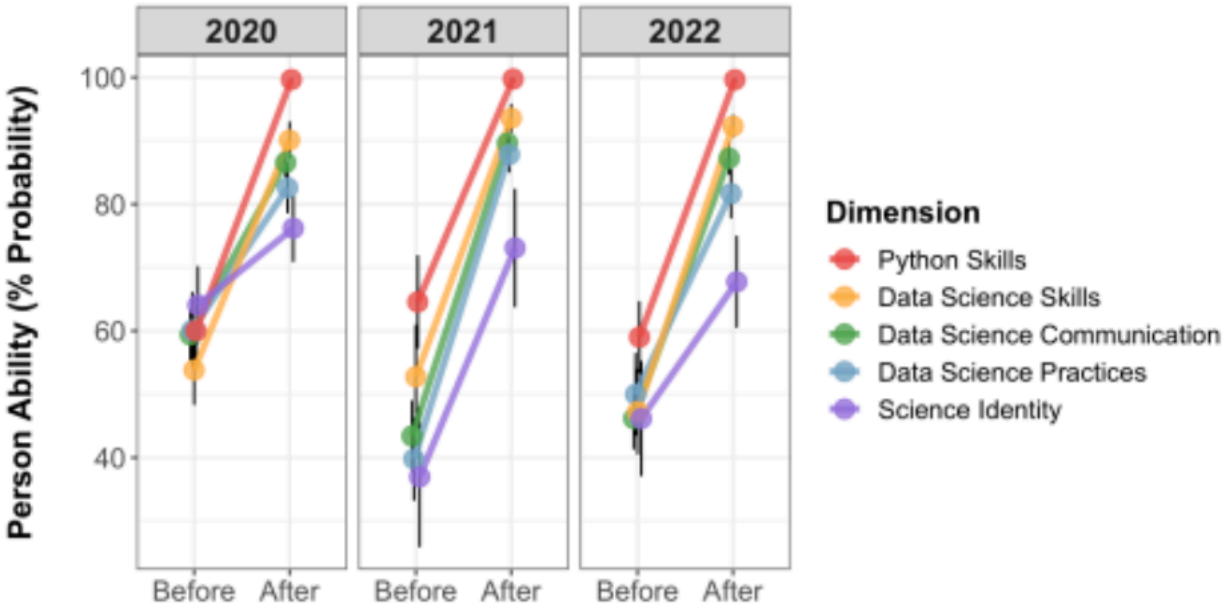

Figure A2: Growth in measures of person ability (% probability) for the five different dimensions across the two timepoints for each of the 3 cohorts of EDSC students (before and after participation in the EDSC program).