Broadening Participation in STEM by Engaging Students in Data Science in Puerto Rico

Steven McGee
The Learning Partnership
mcgee@lponline.net

Willow Kelleigh
The Learning Partnership
wkelleigh@thelearningpartnership.net

Abstract:

This study explores the relationship between students' perceptions of a data science initiative and their expectancy-value beliefs in science. The initiative involved middle and high school students analyzing ecological data from Puerto Rico’s El Yunque rainforest using the Common Online Data Analysis Platform (CODAP). The research is based on Eccles’ Expectancy-Value, which found that expectations of success and the value they place on the field of science is correlated with their future career choices about science. The research analyzed data from 461 students and 20 teachers to examine the relationship between the perceived difficulty of the tasks and teachers’ engagement of scientific inquiry with students’ perceptions of expectancy-value. Results showed that higher student perceptions of task easiness and teachers' use of inquiry methods positively influenced students' post-initiative expectancy and value scores. These findings suggest that appropriately challenging scientific inquiry experiences can enhance students' beliefs in their ability to succeed in science and the value they place on the field, potentially promoting long-term participation in STEM fields. The study highlights the initiative’s potential to broaden participation in science among underrepresented Hispanic students in Puerto Rico, aligning with broader efforts to address STEM disparities.

Acknowledgment

The authors were supported in part by National Science Foundation grant DRL- 2049061. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

Suggest Citation

McGee, S. & Kelleigh, W. (2025, March 13). Broadening Participation in STEM by Engaging Students in Data Science in Puerto Rico [Presentation]. National Association of Research on Science Teaching [online]. https://doi.org/10.51420/Conf.2025.7

1 Introduction

The Expectancy-Value model has been widely used in a variety of areas to predict future choices and behavior. For example, Eccles (2009) conducted a longitudinal study of around 1,000 adolescents from southeastern Michigan. The study began when the adolescents were seniors in high school. Eccles and her colleagues surveyed the students' occupational aspirations as well as dimensions of the Expectancy-Value model as it relates to a wide variety of occupations. They analyzed the relationship between the ratings for various occupational fields and the resulting occupational choices. Over time, the adolescents' eventual job choice was strongly predicted by Expectancy-Value, not only in terms of which occupations were chosen, but also in which choices were not selected. When Expectancy and Value were high for a given field, it was likely the career choice of that adolescent. Conversely, if either expectancy or value were low for a given field, it was unlikely to be the career choice of that adolescent. Eccles' research has shown that people's Expectancy of success develops over time based on experiences with the relevant school subjects and individuals' subjective interpretation of those experiences. Perceptions of value are influenced by a number of social factors, such as enjoyment, perceptions of whether the major will meet personal goals, as well as family, friends, and educators' opinions. Therefore, the Expectancy-Value framework can be used to gauge students' identification with the field of science and track whether educational experiences are expanding the number of students who will likely pursue a career in science.

1.1 Data Jam

In this study, we analyze the relationship between students' participation in the Data Jam program and their perceptions of expectancy-value related to the field of science. Data Jam engages middle-school and high-school students in the process of asking their own questions about environmental phenomena in the El Yunque national rainforest in Puerto Rico (McGee et al., 2024). Students explore, analyze and summarize long-term ecological data as evidence to investigate their own scientific research questions. Successful projects require that students engage in a variety of data practices that underlie the field of science. The data science experience culminates with teachers selecting the best projects from their class. Those students are invited to present their research to their peers and to scientists at a student science symposium at the visitor center in El Yunque rainforest.

To analyze their data, students use an open-source data analysis tool developed by the Concord Consortium called the Common Online Data Analysis Platform (CODAP). It has been developed specifically for students, based on decades of research on how students understand data. The features of CODAP enable students to interrogate data by making it easy to organize, visualize, and see relationships in and between data sets. Students can drag and drop variables on to a graph and access a toolbar with a variety of context sensitive analysis tools. For example, if students graph categorial data, the toolbar will provide buttons for counts and percentages. If students graph numerical data, the toolbar provides buttons to calculate and display numerical analysis tools such as the mean, median, boxplots, and standard deviation.

The project has developed a professional development model that aligns to the five key features of effective professional development outlined in Desimone and Garet (2015)'s empirical framework and includes: (1) sustained duration of more than 20 hours of professional development, (2) a focus on disciplinary content through investigations of long-term data, (3) actively engaging teachers in learning with and reflecting on data, (4) supporting coherence with course goals, and (5) fostering collaboration among teachers. The data science professional development model includes 30 hours of professional development at a summer workshop in which teachers actively engage in the student activities using a "student hat" (Lowell & McNeill, 2020) to develop an understanding of the disciplinary content and data science skills.

During the fall semester, students complete three challenge tasks of increasing complexity. The challenge tasks are administered online and serve a formative assessment purpose. Each challenge involves an investigation question, brief background information, and a link to the necessary data in CODAP. Students demonstrate their data analysis skills by conducting the necessary analyses in CODAP and uploading their graphs. In addition, they submit a written argument that uses the graphs as evidence to respond to the investigation question. The expectations of the challenge tasks provide a vision of what skills teachers need to develop with their students prior to the challenge task administration window.

After each set of challenges are submitted, a random sample of the challenge tasks responses is selected for discussion with teachers. The teachers meet in a cross-school, Virtual Lab Meeting (VLM), with 5-7 teachers to examine the student work using a validated looking at student work protocol (McDonald et al., 2013). The protocol provides a systematic means for teachers to focus on how the work demonstrates what students know, what they still need to learn, and stimulates discussion of how to support student learning in subsequent lessons. These discussion provide teachers with insights on ways to continuously improve their practice.

Data Jam provides students with an authentic experience of the process of scientific inquiry with long-term data about local phenomena in Puerto Rico. The features of Data Jam align with the features of experiences that Eccles' research has found to contribute to increased expectancy-value. In this research, we analyze the relationship students' perceptions of the level of difficulty of the data science experience and their post expectancy-value perceptions. The perceptions of the difficulty of the experience provides a proxy for the extent to which the students perceived that the experience provided an appropriate level of challenge. In addition, we examined the extent to which teachers engaging students in the process of inquiry and the extent to which teachers fully participated in the professional development relate to students' expectancy-value perceptions. We address the following research question:

To what extent does engaging students in the practices of science correlate with positive attitudes towards science?

2 Methods

2.1 Sample Size

Data from a total of 461 students are used in this study, from 20 different teachers. The number of students per teacher ranges from 2 to 104, with a median of 17.5 and a mode of 20.

2.2 Sample Demographics

Most of the students in the sample identify as Hispanic, at 88%. Over half of the sample either declined to specify their race or chose "Other". Because of this, in combination with the multitude of ways many Hispanic people understand race, not all of which align with each other, race was not included in the models. See Rodríguez (2000) for further insight on this topic. Regarding gender, 43.6% of the students identified as male, 54.4% identified as female, and 0.2% preferred not to answer. Other gender options were available for students to select (transgender male, transgender female, nonbinary/ nonconforming, and other) but no students selected them. The grade levels of the students ranged from 6th to 12th, with no students in 9th grade. The distribution of the sample grades is shown in Table 1.

Grade	Pct
6th	10.4%
7th	24.3%
8th	7.4%
10th	8.7%
11th	38.2%
12th	11.1%

Table 1: Sample Grades

2.3 Variables

2.3.1 Dependent Variables

The two variables of interest in this study were expectancy and value, each indicated by 3 measured variables from their portion of the EV (Expectancy Value) statements measured on a 6-point Likert scale, then combined to use as a construct (Barron & Hulleman, 2015). These questions were asked in the pre-survey administered in September 2023 at the beginning of the Data Jam experience as well as the post-survey administered in January 2024 after the Data Jam experience was completed. The post-survey response was the dependent variable, and the pre-survey was an independent variable. They construct and survey items are below:

- **Expectancy**:
  - *I know that I can learn by science class material.*
  - *I believe that I can be successful in my science class.*
  - *I am confident that I can understand the material in my science class.*
- **Value**:
  - *I think my science class is important.*
  - *I value my science class.*
  - *I think my science class is useful.*

For both expectancy and value, the post measurement has been centered across the sample population, while the premeasurement has been centered by teacher. This allows us to see how students' post expectancy and value measures differ while retaining the context of the class that they were in.

2.3.2 Independent Variables

Student level independent variables considered in the models include the following:

- The relevant pre-measurement variable for expectancy or value, as discussed in 2.3.1.
- Whether or not the student identifies as Hispanic, where students who declined to answer were grouped with students who identified themselves as not Hispanic, since Hispanic-identifying students made up most of the sample at 88%. The default for the intercept of the models was Hispanic identifying students.
- The gender of the student, including male, female, and prefer not to answer. The default for the intercept of the models was male identifying students.
- The grade level of the student, ranging from 6th grade to 12th grade, with the notable exception of any students from 9th grade. The default for the intercept of the models was 6th graders.
- The perceived easiness of Data Jam, which were collected on a 7-point Likert scale and then averaged across the different elements (teamwork, choosing a dataset, graphing, analyzing/interpreting data, explaining underlying science, displaying data creatively, writing a report, and timeliness). This was then centered across the sample population.

Teacher level variables considered in the models include the following:

- The average class premeasurement value for either expectancy or value, depending on which is being modeled. This was centered on the average for the sample population.
- A participation metric which ranged from 0 to 1 based on the percentage of challenges a teacher had their class complete and the number of virtual lab meetings the teacher attended. This served as a proxy for the level of participation in the professional development.
- Inquiry-Based Science Inquiry Scale was adapted from an index used by Allensworth, Correa, and Ponisciak (2008). Students' perceptions were collected on a 5-point Likert scale and then averaged across several aspects of inquiry-based science teaching (using lab equipment, writing lab reports, using evidence to support hypotheses, using graphs, creating graphs, and creating models). This was then centered across the sample population.

Independent variables were included on a theoretical basis and checked for multicollinearity as recommended by Kalnins and Praitis Hill (2023). The multicollinearity check showed mild correlation between variables.

3 Results

3.1 Summary Statistics of Measured Variables

Table 2 provides summary statistics for all measured variables.

Question	Mean	SD	Range
Pre-Expectancy	5.07	0.70	1= Strongly Disagree to 6= Strongly Agree
Post Expectancy	5.00	0.81	1= Strongly Disagree to 6= Strongly Agree
Pre-Value	5.25	0.73	1= Strongly Disagree to 6= Strongly Agree
Post Value	5.14	0.83	1= Strongly Disagree to 6= Strongly Agree
Perception of Easiness	3.28	0.62	1= Very Difficult to 5= Very Easy
Use of Inquiry Methods	2.84	0.37	1= Never to 5= Almost Every Day
Teacher Participation	0.65	0.22	0= No Participation to 1= Full Participation

Table 2: Measured Variable Results

3.2 HLM Model

A mixed model, also known as a Hierarchical Linear Model (HLM) approach was used because of the inherent structure of the data, in that two students with the same teacher are less independent from each other than two students from two different teachers. Two sets of models were created, one for expectancy and one for value. Each set of models contains 3 nested models. The first contains only the relevant pre-measurement variable, the second contains all student level variables, and the third contains both the student level variables and the teacher level variables. The model was created in jamovi, using the GAMLj3 package which runs on R (The jamovi project, 2024; R Core Team, 2023; Gallucci, M., 2019; Lüdecke, Ben-Shachar, Patil & Makowski, 2020).

3.3 Visualization of Results

Table 3 provides a visualization of the results of the expectancy models, and Table 4 provides a visualization of the results of the value models. The statistically significant variables ($p<0.05$) are bolded. For both models, the premeasurement of the variable in question was significant, along with the perception of easiness of the data science initiative, and the use of inquiry methods by teachers. The teachers' level of

	Model 1 (Pre Measurement Only)				Model 2 (Student Level Only)				Model 3 (Full Model)			
Variable	Estimate	SE	t(df)	p	Estimate	SE	t(df)	p	Estimate	SE	t(df)	p
Intercept												
Intercept	0.05	0.07	0.79 (18)	0.438	-0.02	0.12	-0.16 (98)	0.871	-0.09	0.23	-0.37 (31)	0.717
Student Level Variables												
Pre Expectancy Student Deviation	**0.48**	**0.05**	**9.55 (395)**	**<.001**	**0.45**	**0.05**	**8.73 (386)**	**<.001**	**0.45**	**0.05**	**8.71 (387)**	**<.001**
Hispanic (non-Hispanic)	/	/	/	/	-0.13	0.12	-1.11 (398)	0.266	-0.13	0.12	-1.12 (396)	0.265
Gender (Female)	/	/	/	/	-0.06	0.07	-0.79 (397)	0.428	-0.06	0.07	-0.91 (396)	0.363
Gender (PNA)	/	/	/	/	-0.15	0.26	-0.57 (388)	0.572	-0.17	0.27	-0.62 (389)	0.535
Grade Level (7th)	/	/	/	/	-0.28	0.15	-1.81 (87)	0.074	-0.25	0.16	-1.57 (106)	0.120
Grade Level (8th)	/	/	/	/	0.00	0.22	0.00 (27)	0.999	-0.19	0.21	-0.90 (33)	0.375
Grade Level (10th)	/	/	/	/	-0.35	0.22	-1.55 (20)	0.136	-0.44	0.17	-1.96 (27)	0.061
Grade Level (11th)	/	/	/	/	-0.13	0.19	-0.71 (19)	0.486	-0.26	0.17	-1.54 (18)	0.141
Grade Level (12th)	/	/	/	/	-0.07	0.23	-0.29 (19)	0.773	-0.36	0.21	-1.70 (16)	0.109
Perception of Data Jam Easiness	/	/	/	/	**0.20**	**0.06**	**3.28 (399)**	**0.001**	**0.20**	**0.06**	**3.28 (396)**	**0.001**
Teacher Level Variables												
Use of Inquiry Methods	/	/	/	/	/	/	/	/	**0.49**	**0.20**	**2.48 (15)**	**0.025**
Teacher Participation	/	/	/	/	/	/	/	/	0.02	0.29	0.07 (15)	0.946
Class Pre Expectancy	/	/	/	/	/	/	/	/	0.42	0.30	1.40 (23)	0.175
Model Fit and Variance												
	Value	df	p		Value	df	p		Value	df	p	
LRT	24.5	1	<.001		9.24	1	0.002		2.88	1	0.090	
AIC	905				911				911			
BIC	921				964				976			
ICC	0.104				0.089				0.043			

Note: Estimate = Unstandardized coefficients, SE = Standard error, LRT = Likelihood ratio test, AIC = Akaike information criterion, BIC = Bayesian information criterion, ICC = Intraclass correlation

Table 3: HLM for Expectancy

	Model 1 (Pre Measurement Only)				Model 2 (Student Level Only)				Model 3 (Full Model)			
Variable	Estimate	SE	t(df)	p	Estimate	SE	t(df)	p	Estimate	SE	t(df)	p
Intercept												
Intercept	0.05	0.07	0.70 (21)	0.489	0.07	0.13	0.53 (83)	0.594	-0.01	0.24	-0.05 (30)	0.965
Student Level Variables												
Pre Value Student Deviation	**0.46**	**0.05**	**8.91 (394)**	**<.001**	**0.44**	**0.05**	**8.34 (356)**	**<.001**	**0.44**	**0.05**	**8.25 (386)**	**<.001**
Hispanic (non-Hispanic)	/	/	/	/	0.07	0.12	0.54 (392)	0.589	0.06	0.12	0.52 (392)	0.128
Gender (Female)	/	/	/	/	0.12	0.07	1.64 (392)	0.101	0.11	0.07	1.53 (392)	0.128
Gender (PNA)	/	/	/	/	-0.01	0.28	-0.02 (385)	0.982	-0.02	0.28	-0.08 (386)	0.936
Grade Level (7th)	/	/	/	/	-0.28	0.17	-1.60 (125)	0.110	-0.26	0.16	-1.62 (83)	0.109
Grade Level (8th)	/	/	/	/	0.01	0.25	0.05 (32)	0.961	-0.11	0.24	-0.45 (47)	0.655
Grade Level (10th)	/	/	/	/	-0.46	0.27	-1.73 (24)	0.096	-0.47	0.25	-1.94 (29)	0.062
Grade Level (11th)	/	/	/	/	-0.15	0.22	-0.66 (23)	0.515	-0.25	0.19	-1.35 (22)	0.188
Grade Level (12th)	/	/	/	/	0.07	0.27	0.25 (24)	0.805	-0.25	0.23	-1.10 (19)	0.283
Perception of Data Jam Easiness	/	/	/	/	**0.16**	**0.06**	**2.58 (394)**	**0.010**	**0.16**	**0.06**	**2.61 (392)**	**0.009**
Teacher Level Variables												
Use of Inquiry Methods	/	/	/	/	/	/	/	/	**0.51**	**0.23**	**2.24 (26)**	**0.034**
Teacher Participation	/	/	/	/	/	/	/	/	0.03	0.30	0.10 (15)	0.920
Class Pre Value	/	/	/	/	/	/	/	/	0.45	0.26	1.69 (19)	0.107
Model Fit and Variance												
	Value	df	p		Value	df	p		Value	df	p	
LRT	37.9	1	<.001		27.2	1	<.001		4.67	1	0.031	
AIC	927				936				935			
BIC	943				988				999			
ICC	0.120				0.136				0.050			

Note: Estimate = Unstandardized coefficients, SE = Standard error, LRT = Likelihood ratio test, AIC = Akaike information criterion, BIC = Bayesian information criterion, ICC = Intraclass correlation

Table 4: HLM for Value

participation in the professional development was not statistically significant. This can be interpreted as below for each variable:

- An increase in the deviation of a student's premeasurement score from their class average is associated with an increase in the deviation of a student's post measurement score from the population average.
- An increase in a student's perception of the easiness of the data science projects is associated with an increase in the deviation of a student's post measurement score from the population average.
- An increase in a teacher's use of inquiry methods of teaching is associated with an increase in the deviation of a student's post measurement score from the population average.

4 Contribution to the Teaching and Learning of Science

Hispanic workers are underrepresented in STEM disciplines (Taylor & Arbeit, 2024) and the opportunities for broadening participation in STEM for Puerto Rican students face acute challenges due to the severe fiscal crisis in Puerto Rico (Melendez-Ackerman & Colón, 2022). Initiatives like Data Jam are a means to address these challenges by providing teachers with cost effective approaches for engaging students in scientific inquiry about their local environment. This research provides evidence that engaging students in the process of inquiry at an appropriate level of challenge, through programs like Data Jam, can positively influence students' perceptions of their expectancies of success and the value that they place in the field of science. Based on Eccles' research, these results provide predictive evidence that Data Jam may be contributing to the long-term broadening of participation in science for students in Puerto Rico.

5 References

Allensworth, E. A., Correa, M., & Ponisciak, S. (2008). From High School to the Future: ACT Preparation–Too Much, Too Late Chicago: Consortium on Chicago School Research at the University of Chicago.

Barron, K., & Hulleman, C. (2015). Expectancy-Value-Cost Model of Motivation. In J. D. Wright (Ed.), International Encyclopedia of Social and Behavioral Sciences, 2nd Edition: Motivational Psychology (Vol. 8, pp. 503-509). https://doi.org/10.1016/B978-0-08-097086-8.26099-6

Eccles, J. S. (2009). Who Am I and What Am I Going to Do With My Life? Personal and Collective Identities as Motivators of Action. *Educational Psychologist, 44*(2), 78-89.

Gallucci, M. (2019). *GAMLj: General analyses for linear models*. [jamovi module]. Retrieved from https://gamlj.github.io/.

Kalnins, A., & Praitis Hill, K. (2023). The VIF Score. What is it Good For? Absolutely Nothing. Organizational Research Methods, 28(1), 58-75.

https://doi.org/10.1177/10944281231216381

Lüdecke, Ben-Shachar, Patil & Makowski (2020). *Extracting, Computing and Exploring the Parameters of Statistical Models using R*. CRAN. link.

Meléndez-Ackerman, E. & Colón, J. L. Colón (2022). Killing the golden goose for STEM. *Science* **375**,1205-1205. https://doi.org/10.1126/science.abo1138

McGee, S., Báez-Rodríguez, N., Castro-Rivera, C., González-Orellana, N., McGee-Tekula, R., & Rubin, A. (2024). Luquillo Data Jam: A powerful model for broadening participation in science. *The Earth Scientist 41*(2), 6-18. https://nestanet.org/TES

R Core Team (2023). *R: A Language and environment for statistical computing*. (Version 4.3) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from CRAN snapshot 2024-01-09).

Rodríguez, C. E. (2000). Changing Race: Latinos, the Census and the History of Ethnicity. United Kingdom: NYU Press.

Taylor, D. & Arbeit, C.A. (2024, May). The STEM Labor Force: Scientists, Engineers, and Skilled Technical Workers [Report]. Washington, DC: National Science Foundation.

The jamovi project (2024). *jamovi*. (Version 2.5) [Computer Software]. Retrieved from https://www.jamovi.org.