FIRST® 10 Year Longitudinal Study: Study Methodologies Final Year of Study

(July 2024)

The *FIRST* Longitudinal Study was designed and implemented in 2012 to provide a rigorous assessment of the short and longer- term impacts on three of *FIRST*'s major programs – the *FIRST®* LEGO® League (FLL®), the *FIRST®* Tech Challenge (FTC®), and the *FIRST®* Robotics Competition (FRC®) – on the educational and career trajectories of the programs' participants. The goal of the study is to determine whether, as a result of participation in *FIRST*, middle and high school-aged young people are more likely to gain and sustain an interest in STEM, pursue STEM-related education in high school and college, and take steps towards ultimately entering into STEM-related careers than are similar youth who do not participate in the program. Other key outcomes at implementation included the development of a variety of attitudes and skills related to success in the 21st Century workplace, including teamwork, problem- solving and communications skills, leadership and service, and the ability to work with others (including competitors).[1] Three major questions guided the study:

(1) **What are the short and longer-term impacts of the FLL, FTC, and FRC programs on program participants?** Specifically, what are the program impacts on a core set of participant outcomes that includes: interest in STEM and STEM-related careers, college-going and completion, pursuit of STEM-related college majors and careers, and development of 21st century personal and workplace-related skills?

(2) **What is the relationship between program experience and impact?** To what extent are differences in program experience – such as time in the program, participation in multiple programs, role on the team, access to Mentors, quality of the program experience – associated with differences in program outcomes? What can we learn about "what works" to guide program improvement?

(3) **To what extent are there differences in experiences and impacts among key subpopulations of *FIRST* participants?** In particular, are there differences in impacts among young women, urban/rural, and youth from low-income communities? If there are differences, what can we learn about why those differences occur and how to address them in the future?

In its final year (2023-2024), the *FIRST* Longitudinal Study expanded on its core data collection through the annual participant and comparison group surveys by adding a qualitative interview study of 42 *FIRST* female participants and analysis using national restricted data from the National Center for Education Statistics. Below we describe the methodology for each of these components implemented in the final year of the study.

[1] The 21st Century Workplace questions were discontinued in the 5th year of follow-up data collection as there were no significant differences found between the two groups in the study in any of these years.

(1) The Annual Follow-up Study Design

The *FIRST* Longitudinal Study was designed to address these questions and provide a rigorous assessment of *FIRST's* short and longer-term impacts by applying both a longitudinal approach, tracking participants in FLL, FTC, and FRC over a period of five or more years, and by incorporating a comparison group into the design. The quasi-experimental, comparison group design is intended to provide an answer to the question "What would have happened in the absence of *FIRST*?" by comparing the changes in attitudes and educational and career trajectories of program participants with those of youth with similar interests at baseline who do not participate in *FIRST*.

To accomplish this, the Longitudinal Study began tracking approximately 1,273 students (822 *FIRST* participants and 451 comparison students) over ten years beginning with entry of the *FIRST* participants into the program.[2] The study is focused on new participants in *FIRST* (i.e., those with no prior participation in the FLL, FTC, or FRC program) so that it can track team members from their point of entry into the program. Team members were recruited to the study from a nationally representative sample of over 200 "veteran" *FIRST* teams in 10 states. Comparison group students were recruited from math and science classes in the same schools and organizations where the *FIRST* teams were located. Participant recruitment took place in two waves, with recruitment of an initial group of students in Fall 2012 and recruitment of additional participants in Fall 2013 to increase the size of the overall sample for the study. In the final year of the study, 922 students (72%) completed the survey, including 551 *FIRST* participants (67% of baseline) and 371 comparison students (82% of baseline).

One of the key decisions in designing the study was to employ a comparison group design. In impact studies like these, the preferred evaluation design is often a randomized control trial in which participants are recruited to the program being studied and are then randomly assigned to either the program in which they participate or to a control/comparison group where they are excluded from program participation. The randomization process is intended to ensure that program participants and comparison students have similar baseline interests and characteristics and to control for any inherent bias in the sample in terms of those who normally join or not join the program. For *FIRST*, however, a randomized control model was determined to not be feasible, for several reasons. In general, local *FIRST* teams recruit as many team members as they can, so it was unlikely that there would be sufficient additional applicants for randomization. Moreover, since teams tend to accept all interested youth, it would have been viewed as unethical to actively exclude interested young people from the program or to prohibit them from joining a *FIRST* team during the extended period of the study.
Consequently, the decision was made early in the design process to pursue a "quasi-experimental," comparison group design that would recruit non-participating students to serve as the comparison group for the analysis. The decision to recruit comparison students from math and science classes in the schools and organizations where *FIRST* teams were located was an effort to recruit a comparison group that would include at least some students with

[2] The study includes 206 team members from FLL teams, 248 from FTC teams and 368 team members from FRC teams. The comparison group includes 195 students in 5th-8th grades and 256 high school (9th-12th grades) students.

substantial interest in STEM, while also controlling for differences at the school or community level.[3]

Data Collection

The primary source for the study is a series of baseline, post-program, and annual follow up surveys of team members and comparison students. A baseline survey of parents provides additional background information on the family context for team members and comparison students, and Team Leader surveys at the end of the first year of team involvement in the study provide additional contextual data on the *FIRST* teams. Surveys have been supplemented by telephone interviews and focus groups with participants in several years of the study.

Baseline surveys were administered to program participants and comparison students as paper-based surveys when they entered the study in Fall 2012 and 2013. Follow-up surveys have been administered as an online survey in each subsequent spring. Exhibit 1 shows the survey response rates for the study throughout the entire study period.

Exhibit 1: Number of Responses Baseline through 10 Year Follow-Up

Group	***FIRST* Participants**		**Comparison Group**		**Total Responses**	
Data Collection Wave	Number	Year to Year Response Rate	Number	Year to Year Response Rate	Number	Year to Year Response Rate
Baseline	822		451		1273	
12-Month Follow-Up (Post-Program)	677	82%	259*		936*	74%*
24-Month Follow-Up	665	98%	411	91%	1076	115%
36-Month Follow-Up	636	96%	409	100%	1045	97%
48-Month Follow-Up	611	96%	406	99%	1017	97%
60-Month Follow-Up	602	99%	397	98%	999	98%
72-Month Follow-Up	550	91%	386	97%	936	94%
84-Month Follow-Up	554	101%	389	101%	943	101%
96-Month Follow-Up	570	103%	385	99%	955	101%
108-Month Follow-Up	559	98%	379	98%	938	98%
120-Month Follow-Up	551	99%	371	98%	922	98%

[3] The decisions concerning comparison group design were among the most challenging for the study. Generally, *FIRST* team members, as participants in an after-school program, self-select into the program, and it was anticipated that a substantial percentage would enter the program with a prior interest in STEM. Brandeis staff and the advisory groups for the study explored a variety of options including recruiting students from other after-school programs; having team members identify non-participating friends; and recruiting students from other, non-*FIRST* schools. Ultimately, it was decided to recruit comparison students from math and science classes at the schools where *FIRST* teams were located as the most feasible, most likely to control for school-level effects, and as the most likely to result in recruiting sufficient numbers of comparison students for the study.

Response Rate Baseline to 10 Year Follow-up	**67%**		**82%**		**72%**	

**The initial group of comparison students did not complete a post-program survey but have participated in all subsequent follow-up surveys.*

Study Outcomes

The major focus of the study is on *FIRST*'s impacts on STEM-related interests, attitudes, and behaviors. Key outcomes, developed in collaboration with staff at *FIRST* and with the program and technical advisory groups during the planning phase of the study, include a combination of interest and attitudinal measures (for example, increased interest in STEM and STEM-related careers, sense of educational efficacy, and postsecondary aspirations); measures of self-reported life and workplace skills; and shorter and longer-term behavioral measures such as increased STEM-related course-taking, postsecondary STEM course-taking and college majors, and continued involvement in *FIRST*. Exhibit 2 provides an overview of the key outcome measures.

Exhibit 2: Key Outcome Measures in the Final Year of the Study (Year 10)

STEM-Related Interest and Attitude Scales	Behavioral Measures
• STEM Interest (Level of interest in science, technology, engineering and mathematics) • STEM Activity (involvement in non-school STEM activities) • STEM Careers (interest in STEM-related careers, such as scientist, engineer, computer specialist, etc.) • STEM Identity (extent to which students see themselves as science, math or technology people) • STEM Knowledge/ Understanding (awareness of applications of STEM in real world, interest in learning more about STEM).	• Interest in STEM Majors in College/Declared Majors • STEM-Related College Course- taking • Involvement in College STEM- Activities (Clubs, competitions, internships, summer jobs) • STEM-related College Grants and Scholarships • Employment in STEM

In addition to the key outcome measures, the baseline surveys collected demographic information including age, gender, race/ethnicity, ESL status, and grade in school as well as information on program participation and academic background (grade point average, honors courses at baseline). Parent surveys provided information on family income and parental support for their children's involvement in STEM. As discussed below, these baseline characteristics were used in the analysis to control for differences between participants and comparison group member characteristics at baseline and to control for the influence of characteristics like race or gender on outcomes. The survey items were drawn from a mix of existing national surveys (for example, the U.S. Department of Education's National High School Longitudinal Study of 2009), questions that had been used in previous evaluation studies, and items developed specifically for this study. The surveys were piloted with students on local after school robotics teams and revised based on their feedback. A summary of the scale measures used in the study can be found at the end of this document.

Approach to Assessing Impacts

The basic method for assessing impact in this study is by comparing outcomes for participants and comparison students while controlling for differences between the two groups at baseline. As shown in Exhibit 1, the current analysis (based on 10 year data) includes data from eleven rounds of participant surveys, including Baseline survey data, Post-Program (end of the first year) data for most study participants, and nine annual follow-up surveys.

To make full use of the multiple data points that are available, the study uses a "repeated measures linear mixed models" method of analysis as the primary method of statistical analysis. The "Mixed" method is a form of multivariate analysis that allows the inclusion of covariates (control variables) to control for differences in participant characteristics and settings in the analysis and for the use of repeated measures (i.e., multiple data points) over time. The mixed methods approach, unlike many other statistical tests, also allows the use of all of the data available in developing estimates of the outcomes, even when there is missing data for some students in the sample at some of the data points. [4]As a result, the mixed methods approach makes it possible to use data from all five sets of surveys despite the fact that not all students completed every one of the surveys.

The mixed methods analyses provide estimated outcome measures for team members and comparison students that take into account the various control measures and differences at baseline. When compared, the differences in those outcomes provide the measure of impact from the program – whether there are statistically significant differences in the gains (or declines) for *FIRST* team members and comparison students. For this study, adjustments for differences between the participant and comparison groups at baseline include covariates for

[4] For background of the mixed-models method, see A.A O'Connell and D.B. McCoach, eds. (2008). *Multilevel Modeling of Educational Data*. Charlotte, NC: Information Age Publishing; and J.D. Singer (1998). "Using SAS PROC MIXED to Fit Multi-Level Models, Hierarchical Models, and Individual Growth Models." *Journal of Educational and Behavioral Statistics*, 24(4), pp. 323-355.

gender, race/ethnicity, family income, participation in STEM honors courses at baseline, and baseline parental support for STEM. Analysis of behavioral measures (e.g., college major, college course-taking) also includes STEM interest at baseline as a covariate.

The study also uses a second type of multivariate analysis: Logistic Regression Analysis or "Logit." Logit analysis estimates the relative probability that *FIRST* participants and comparison students will achieve a particular outcome, taking into account differences between the groups at baseline. In this study Logit analysis is used to assess whether *FIRST* participants are significantly more (or less) likely than comparison students to show an increase from baseline to follow-up on the various scale score measures (such as STEM interest); Logit is also used to examine whether *FIRST* participants are significantly more likely to want to major in engineering or take engineering courses. The "odds ratio" produced by the Logit analysis is a measure of the relatively likelihood that one group or another will achieve that particular outcome (for example, that "*FIRST* participants are 3.0 times more likely to show a gain in STEM interest" or 3.1 times more likely to want to major in engineering) after taking into account differences at baseline. As with the "Mixed" analysis, the Logit analyses in this study include covariates for gender, race/ethnicity, family income, participation in STEM honors courses at baseline, and baseline parental support for STEM and, when appropriate, STEM interest at baseline.

In sum, the two methods provide two ways of looking at program impacts. The "Mixed" analysis basically looks at the *difference in average gains* (or declines) between the two groups in the study; the Logit analysis determines whether, on average, one group or the other was significantly more likely to show *any* gain from baseline to follow-up. It is important to note that in some cases, *FIRST* participants and comparison students are equally likely to show a gain on a particular measure (no significant difference using the Logit analysis), but that on average, the gains that do take place for *FIRST* participants are significantly greater than those for comparison students (positive, statistically significant impacts using the "Mixed" analysis). Both results are accurate and appropriate – they provide two somewhat different perspectives on impact (average gain vs. likelihood of gain).[5]

Comparison Group

A critical part of the analysis of program impacts is the use of a comparison group to estimate what would have happened in the absence of the program. As noted earlier, the comparison group for the study is comprised of non-participating students (i.e., students not involved in *FIRST*) who were recruited into the study through math and science classes at the schools and organizations where the *FIRST* teams in the study are located. The goal of that effort was to recruit a comparison group that would include at least some students with substantial interest

[5] The Logit analysis differs from the "Mixed" approach in one other important respect – it only makes use of data from two points in time, in this case the baseline and 1 Year Follow-up survey. Consequently, the sample sizes for the Logit analysis are substantially smaller than for the "Mixed" analysis, making it less likely for results to show statistical significance than in the "Mixed" analysis, even when differences are quite large. As a result, the study restricts the use of the Logit analysis to the analysis of impacts for the sample as a whole and the analysis of impacts by program and did not use Logit in the analysis of other subgroup differences.

in STEM, while also controlling for differences at the school or community level. Approximately 450 students were recruited into the comparison group over the two years of recruitment for the study. Comparison group students have been told that they are participating in a study of STEM-related interests and activities (the SciTech study) and, as a result, are often referred to as "SciTech" students in the study reports.

Exhibit 3 provides an overview of the baseline characteristics of *FIRST* team members and comparison students in the study. As the table shows, the comparison group students and participants are relatively well-matched on some measures and show statistically significant differences on other. In general, the two groups are similar (i.e., no significant differences) in terms of their average age, ethnic background (percent Hispanic), the types of communities they live in, their academic performance (grades) and their educational aspirations. They also tend to come from families with similar socioeconomic backgrounds – parental education and family income. The two groups differ in the mix of middle and high school students (more *FIRST* students were in 9th-12th grade at baseline), the percentage of White students and youth of color (*FIRST* has a much higher percentage of Asian students; the comparison group has a substantially higher percentage of White students), the percentage of students for whom English was their first language (lower in *FIRST*), and the proportions attending different types of schools. While statistically significant (i.e., not likely to be random differences), the differences are generally not large and can be controlled for in the analysis.

Not surprisingly, there are significant differences between the groups at baseline on a number of measures of STEM interest and attitudes for both students and their parents. In terms of family environment, parents of *FIRST* participants are significantly more likely to have been employed in a STEM-related field, to consider it important that their child participate in STEM-related activities, and to encourage their child to pursue STEM interests and careers. *FIRST* participants also score significantly higher at baseline on the measures of STEM interests and attitudes used in the study. It is important to note that, while *FIRST* participants clearly enter the program (and study) with higher levels of interest in STEM, there are no significant differences on most of the baseline scale measures for the non-STEM outcomes, including academic self-concept, college support, Self-Efficacy, and self-assessed 21st Century Skills.

These differences form an important context for the study: a key goal of the analysis is to control for these baseline differences so that the participants and comparison group students are as comparable as possible. As noted above, the analysis is designed to control for these differences in two ways. First, both the mixed methods and logit approaches take baseline measures into account in calculating outcomes. In that regard, baseline differences on core outcome measures are controlled for as part of the calculation of the outcome estimates. In addition, the models used for developing the impact estimates include a number of covariates (control variables) that provide an additional adjustment for differences between participant and comparison students in the sample. As noted earlier, the final models used for the impact analyses in the study include covariates for gender, race (Asian, White, Black), socioeconomic status (income), parental support for STEM, and baseline involvement in STEM (more honors or advanced STEM-related courses at baseline and, where possible, baseline STEM interest.[5]

Summary

The *FIRST* Longitudinal Study represents an effort to address a core set of questions about the impact of participation in *FIRST* through as rigorous an analysis as possible, given the practical constraints on the overall research design. The students participating in the study are broadly representative of the range of students participating in *FIRST* programs. Comparison students were recruited with the goal of including students with similar demographic characteristics, levels of academic achievement, and interest in STEM. The measures used in the study reflect key outcomes for *FIRST*, were developed in collaboration with *FIRST* staff and advisors and draw on established assessment tools. The longitudinal data collected through the annual surveys not only makes it possible to address longer-term outcomes of program participation, but to examine patterns of participation over time. Finally, the analysis methods are designed to make effective use of the data and to control for baseline differences between participants and comparison students.

[5] Most of the direct measures of STEM interest, including the STEM interest scale, could not be used as control variables since they were included as outcomes in the analysis. Several additional variables were included in the model in the early analyses, including community type (urban/rural/ suburban), parent's education (at least one parent with a BA), and ESL status (English as a primary language). These variables were ultimately dropped from the model when it was found that they were consistently non-significant as predictors in the analysis.

Exhibit 3: Participant and Comparison Group Characteristics at Baseline

Measure	FIRST	SCITECH	ALL
Gender*			
Male	67.8%	41.5%	58.5%
Female	32.2%	58.5%	41.5%
Average Age	13.96	14.14	14.02
School Level*			
5th-8th Grade	28.5%	41.5%	33.1%
9th – 12th Grade	66.7%	56.8%	63.2%
Other	4.8%	1.8%	3.8%
Race/Ethnicity*			
Asian	17.9%	10.2%	15.2%
Black/African-American	8.5%	6.6%	7.8%
White	67.8%	82.9%	73.0%
Ethnicity (NS)			
Hispanic	16.0%	10.0%	14.5%
Other Demographic Characteristics			
ESL (English as first language)*	79.3%	85.5%	81.5%
US Born (NS)	90.3%	93.0%	91.3%
Special Education (NS)	8.1%	3.3%	7.3%
Geography (NS)			
Urban	26.0%	23.2%	25.0%
Suburban	51.3%	53.0%	51.9%
Rural	22.7%	23.9%	23.1%
School Type*			
Regular Public School	71.3%	75.1%	72.6%
Charter School	3.7%	.5%	2.6%
Magnet School	15.3%	7.3%	12.5%
Private School	7.4%	15.6%	10.3%
Academic Performance - Grades (NS)			
Mostly A's	49.5%	49.4%	49.5%
A's and B's	34.0%	36.4%	34.9%
Student's Educational Aspirations (NS)			
BA Degree or More	95.2%	96.4%	95.7%
Parent's Education (Highest Degree) (NS)			
BA Degree or More	59.4%	58.6%	59.1%
Family Income (NS)			
Under $50,000	26.9%	21.7%	25.2%
$50,000- $100,000	32.5%	34.8%	33.2%
$100,000 and over	40.5%	43.5%	41.6%
Parent Employment/Experience in STEM*			
At least 1 Parent ever employed as engineer, scientist, programmer or other STEM field.	49.3%	40.8%	46.3%

Measure	FIRST	SCITECH	ALL
Parent Support for STEM*			
Importance of having child participate in STEM activities (Important/Very Important)*	91.5%	75.4%	86.0%
Parent Encouragement of STEM (5 pt. scale)*	4.2	3.9	4.1
Parent encouragement of STEM careers (7 pt. scale)*	5.4	4.7	5.2
Participant Baseline Scale Scores	**FIRST**	**SCITECH**	
Survey Scales(average baseline scale score)			
STEM Interest*	4.1	3.7	
STEM Activity*	3.4	3.1	
STEM Careers*	4.5	3.7	
STEM Identity*	3.1	2.9	
STEM Knowledge*	5.6	4.9	
Academic Self-Concept	5.71	5.71	
College Support	2.18	2.21	
Self-Efficacy/Prosocial	5.5	5.5	
21st Century Skills	3.1	3.2	
Teamwork/Collaboration subscale	3.3	3.4	
Problem-solving subscale	3.1	3.1	
Communications subscale	2.9	3.0	

Note: An asterisk (*) indicates differences between participants and comparison group members that are statistically significant at p≤ .05. (NS) stands for not significant.

Survey Scale Sources

Domain	Source	Items
Interest in STEM	Brandeis University. Developed for *FIRST* Longitudinal Study (FLS) Alpha = .67	How interested are you in science, technology, engineering and/or math (STEM)? Please mark on a scale from 1 (Not interested) to 5 (Very interested). a. Science b. Technology c. Engineering d. Math
Involvement in STEM activities	Adapted from US Department of Education, High School Longitudinal Study of 2009 (Items c-f added). Alpha = .76	Other than for school, how much do you like to do the following? Please mark on a scale from 1 (Do not like at all) to 5 (Like a lot). a. Read science books and magazines? b. Visit web sites for information on computers and technology? c. Talk with friends or family about science and technology? d. Watch programs on science and technology on television (for example: Science Channel, National Geographic, Discovery Channel)? e. Design web pages? f. Take apart things (like motors, computers, toasters) to see how they work?
	Adapted from US Department of Education, High School Longitudinal Study of 2009. (Items c-f added)	Last school year [year], which of the following types of activities did you participate in through a club, camp, or a competition, in school or out of school? (Mark all that apply.) Do not include participation in *FIRST*. a. Math b. General Science (Biology, physics, chemistry, etc.) c. Robotics d. Computer/ technology e. Engineering f. Environment (clean up clubs, etc.)

Domain	Source	Items
Interest in STEM careers	Adapted from Barker, 4-H Robotics and GPS/GIS Interest Questionnaire (items e-g added). Alpha = .81	How interested are you in each of the following jobs related to STEM (science, technology, engineering, and mathematics)? Please mark one response in each row using the scale from 1 (Not interested at all) to 7 (Very interested). If you are not sure, please give us your best answer. a. Scientist b. Engineer c. Mathematician d. Computer or Technology Specialist e. STEM Educator/ Teacher f. Inventor g. Skilled technician (for example: auto or aircraft mechanic, machinist, electrician, construction)
STEM identity	Adapted from US Department of Education, High School Longitudinal Study of 2009 (Items i-l added) Alpha = .70	Now we are going to ask you a few questions about your beliefs about math and science. How much do you agree or disagree with the following? (Four point scale. Responses include: Strongly Disagree, Disagree, Agree, Strongly Agree) a. I see myself as a math person. b. Others see me as a math person. c. Most people can learn to be good at math. d. You have to be born with the ability to be good at math. e. I see myself as a science person. f. Others see me as a science person. g. Most people can learn to be good at science. h. You have to be born with the ability to be good at science. i. I see myself as a technology person. j. Others see me as a technology person. k. Most people can learn to be good at technology. l. You have to be born with the ability to be good at technology.

Domain	Source	Items
Understanding of STEM	Center for Youth and Communities, Brandeis University, adapted from prior *FIRST* evaluation studies. Alpha = .94	We are interested in learning about how you think about yourself and your future. Using a scale from 1 (Not True at All for Me) to 7 (Very True For Me), please tell us how true each of the following statements are for you. a. I want to learn more about science and technology. b. I can use math and science to do something interesting. c. I have a good idea of what I want to study in college or technical school. d. I am interested in having a job or career that uses science and technology. e. I understand different ways that science and technology can be used to solve problems in the real world. f. I have a good understanding of how engineers work to solve problems. g. I know about a variety of jobs and careers in STEM (science, technology, engineering and/or mathematics). h. I have the kinds of skills that are needed to be a scientist or engineer. i. I can make a good living as a scientist or an engineer. j. I would enjoy working as a scientist or an engineer. k. I can use math and science to make a difference in the world.

Domain	Source	Items
Quality of Program Experience	Center for Youth and Communities, Brandeis University. Adapted from prior *FIRST* evaluation studies. Based on elements of effective youth program in Eccles and Grootman, Eds (2002) Alpha = .814	How well do the following statements describe your experience on your FIRST Robotics team this year? For each statement, please tell us whether you strongly agree, agree, disagree or strongly disagree. a. Students on my team made the important decisions, not the adults. b. I had a chance to do lots of different jobs on my team. c. I had important responsibilities on my team. d. I had a chance to play a leadership role on my team. e. My team learned how to work well together. f. My team really listened to my ideas. g. The adults on my team did most of the difficult jobs in building the robot. h. I had a chance to get to know at least one of the adults on my team very well. i. I felt like I learned a lot from the adults on my team. j. I had a chance to learn about careers in science and engineering on my team. k. I learned about the FIRST college scholarships available to FTC/FRC team members. l. I learned about the importance of Gracious Professionalism. m. I had fun working on my FIRST team. n. I felt like I really belonged on my team. o. I almost always felt that my team had a good chance to win something at the FIRST competition. p. I felt I was an important part of my team. q. The adults on my team helped me think about college and careers

Note: All alpha scores with the exception of the Quality of Program Experience Scale based on Wave 1 and Wave 2 baseline survey data, N=1270. Alpha for Quality of Program Experience calculated based on Wave 1 post-program survey, N=386.

Domain	Source	Items
Mentor Scale		a. *How much did this person help you do any of the following? -Think about the kinds of things I need to study if I want to become a scientist or engineer* b. *How much did this person help you do any of the following? -Learn about science and technology careers* c. *How much did this person help you do any of the following? -Solve a problem with building or programming my team's robot* d. *How much did this person help you do any of the following? -Make me feel like there is someone I could talk with about school or careers*

Measures of Program Participation

There are seven questions that ask what specific role the respondent had on their *FIRST* team. Based on each question, seven various domains are asked about; Designing, Building, Programming, Reviewing Rules, Raising Money, Deciding on Mission/ Strategy, and Operating the robot.

To see which elements were strongly associated with each other, Principal Component Analysis (PCA) was run on the *FIRST* post survey (after their initial baseline year). PCA is a procedure which determines which combinations of variables best reduces the total variance across all measures. Each grouping of variables has an associated eigenvalue. An eigenvalue is a measure of how strong that group is in explaining the total variance. As a rule of thumb, only groupings with an eigenvalue over 1 should be considered.

After running the PCA, we reduced the total number of domains from 7 to 3 PCA groups. These being “building”, “Programming”, and “Team Support”. The chart below explains which questions feed into which PCA groupings. There are two questions that load on to “Building”, one question that loads on to “Programming” and two questions that load on to “Team Support”. There are two questions that do not strongly load on to any specific PCA group.

Because respondents can be on a *FIRST* team for many years and because their role on the team may have changed, we decided to use the average PCA score for each grouping over the first four years of the *FIRST* experience. The same PCA was run for each person’s potentially second, third, and fourth years. The same 3 PCA groups existed across these four years. If a person had only 1 year of FIRST, that was the PCA used. If they had 2 years of FIRST, an average PCA was used over those 2 years. If they had 3 years, an average over 3 years was used. If they had 4 years, an average over 4 years was used.

Domain name:	**Question:**	**PCA group:**
Design	Designing the team's robot or a specific part of the robot	Building
Build	Building the robot or a specific part of the robot	Building
Programming	Programming the robot	Programming
Rules	Reviewing competition rules/gathering information for the team	Team Support
Money	Raising money or doing publicity for the team	Team Support
Mission	Deciding on the team's overall mission/strategy for the competition	
Operate	Working on/setting up or operating the robot at a tournament	

(2) Qualitative Interview Study with *FIRST* Female Alumni

The data used in this analysis are derived from the qualitative portion of a larger mixed-method longitudinal study evaluating the *FIRST* program. The longitudinal study has explored the effectiveness of an after-school robotics program on increasing STEM interest and attitudes of children, and on encouraging students to pursue STEM-related education and career options (Meschede et al., 2022). As part of the study, this qualitative project sought to study the impacts of *FIRST* on the academic and career decisions and trajectories of female participants of the program, as well as why *FIRST* has shown greater impacts upon females in comparison to males. To explore these areas, focus groups and interviews were held with *FIRST* alumni who identify as female and who were current participants of the larger *FIRST* longitudinal study. The focus groups and interviews focused on topics including STEM coursework in high school and college, the fields females were pursuing in their current careers, their experiences in the *FIRST* program they attended, and how they believed *FIRST* influenced their post-*FIRST* academic and work experiences. As interview and focus group participants were at various stages in life spanning from college to post-higher education career experiences, this study focuses on responses to questions about all relevant post-*FIRST* experiences.

Current female-identifying participants of the *FIRST* longitudinal study were solicited for interviews to participate in the qualitative study. Individuals were contacted first by a survey introducing the study, asking for confirmation of their interest in participating. Once responses were collected from all survey respondents, a new survey was sent to those expressing interest in the study to schedule or conduct their interview. In an attempt to gather the widest range of former *FIRST* female participants possible – including diversity of race, ethnicity, socioeconomic status, and geographical location, in addition to whether participants were currently involved in STEM fields or not – reminder emails were periodically sent to survey recipients requesting their responses before closing this phase of study recruitment. This process of quota sampling wherein we divided our sample into different strata (Mason, 2002) increased our likelihood of increasing the racial and ethnic diversity of the research sample, allowing us to gather data from groups historically underrepresented in STEM fields. Additionally, including a stratum pertaining to one's current academic or career field allowed us to draw comparisons between women who were and were not participating in STEM at the time of interviews. While this method of convenience sampling involves self-selection which can impact the representation of individuals in a study sample (Robinson, 2014), we found it to be the best option for recruiting a large group of participants who would be willing to share their experiences with us. Our combined usage of both quota and convenience sampling ultimately led to a diverse group of females in our final data sample (see Table 1).

In comparison to the sample of *FIRST* female respondents to the larger longitudinal study's most recent survey, which was distributed in the tenth year of the evaluation, the study sample had some notable differences. In particular, we saw more even distribution of white and non-white participants (50% of each group); and also had greater representation of Black or African American women in our study sample. We did, however, have lower representation of Hispanic females, and were unable to successfully recruit any women of indigenous backgrounds. Note that percentages in Table 1 do not always add up to 100% as demographic data was missing for some participants, due to the baseline survey questions requesting this information being optional.

Exhibit 3: Interview Sample in comparison to Longitudinal Study Sample of *FIRST* Female Respondents to Year 10 Survey.

Demographics	**Study Sample (N=42)**	**Sample of *FIRST* Female Longitudinal Study Respondents to Year 10 Survey (N=186)**
Race		
American Indian or Native Hawaiian	0% (0)	1.8% (3)
Asian	23.8% (10)	26.7% (46)
Black or African American	11.9% (5)	8.7% (15)
White	50.0% (21)	58.1% (100)
Multi-Racial	4.8% (2)	4.7% (8)
Ethnicity		
Hispanic	9.5% (4)	15.1% (28)
Not Hispanic	81.0% (34)	84.8% (156)
Socioeconomic Status		
High-income	64.7% (22)	69.1% (112)
Low-income	35.3% (12)	30.9% (50)
Geographical Location		
Urban	19.0% (8)	26.5% (48)
Suburban	59.5% (25)	54.7% (99)
Rural	9.5% (4)	18.8% (34)
STEM Involvement		
Currently STEM-involved (robotics, engineering, computer science)	31.0% (13)	21.5% (40)
Currently STEM-involved (all other science fields)	31.0% (13)	21.5% (40)
Not currently STEM-involved	38.1% (16)	57.0% (106)

Determining the STEM status of interview participants was an important step prior to the analysis process, as doing so helped identify trends across various groups. The categories developed by the research team included the following: STEM-involved in robotics, engineering or computer science; STEM-involved in all other science fields; and non-STEM involved. By distinguishing those who are specifically involved in the STEM fields that *FIRST* teaches to program participants, we were able to identify the ways in which these particular individuals have navigated their educational and professional experiences. Further, as the interview data would later show, the ways in which females defined STEM themselves varied, with some self-identifying as engaging in STEM while working in careers traditionally not considered within these domains. Most notably, several interview participants spoke about their careers as educators who teach STEM fields, which informed how they labeled themselves as working in STEM. Thus, we categorized study participants accordingly in order to highlight how the definition of STEM itself has evolved over time, and how many individuals feel as though their current roles reflect STEM engagement.

In preparation of interviews, the research team formulated interview protocol to guide sessions with study participants. The protocol introduced the study in more detail and confirmed participant consent and their willingness to have sessions recorded. Participants received consent forms prior to interviews and before the research team facilitated a conversation amongst interview attendees asking a series of questions regarding a variety of topics. Interview questions included asking about general thoughts and experiences with STEM, and for those who were not engaged in STEM at the time of the interview, their experiences in their current field. Questions in this domain included the types of STEM courses participants were most interested in during school and why, what their initial experience in these classes was like, and what obstacles or reasons might have led to any hesitancy towards any particular STEM topics. Interviews then asked about participants' experiences in *FIRST*, including how and why they initially joined the program, the composition of their team – whether it was a mixed-gender or all-female group – and who their coaches and mentors were, and what they found to be most exciting about the program. Interview questions also covered how *FIRST* may or may not connect with one's current school or work experiences. Finally, we asked a series of questions to help us understand the trajectories of study participants who were currently involved in STEM fields. Questions included what these women found to be most compelling about their field of study or work and why it was that they have remained involved in it, if and how their pathway after *FIRST* has always involved STEM, and what they believed it took to be successful in STEM. We also asked participants if they believed particular STEM fields were more accessible to women than others, how they believed female representation could be increased in these fields, and what their thoughts were around how males could be allies for women in STEM. In addition to these core questions, we developed a broad range of probing questions, to follow up on how interview participants responded throughout the sessions.

To answer the research questions developed for our interview protocol, we used transcripts from 24 completed small focus groups and interviews (15 focus groups ranging in size from two to three individuals, and 9 interviews), including 42 individuals in total. These sessions were held between April 14th, 2023 through July 14th, 2023. Three pilot interviews were conducted during April, in order to assess the efficacy of the initial interview questions devised by the research team. Doing so assisted in our revision and finalizing of the questions to be asked at the remainder of the interviews. All focus groups and interviews were conducted virtually over Zoom, involved between one to three study participants in each session, and lasted approximately one hour. Sessions followed a semi-structured format that was meant to explore study participants' various academic and career experiences post-*FIRST*. Interviewers used an interview script including probing questions that would help follow all leads. Due to this structure, not all interviews discussed the exact same questions and material.

As we requested permission to record sessions from all study participants, nearly all focus groups and interviews were recorded, with only one pilot interview not having been recorded. The recordings were automatically transcribed through the Zoom platform, and were later cleaned using Otter.ai to ensure accuracy of all language captured. Transcripts were then imported into Atlas.ti, a software package that enables computer-based qualitative data analysis. Each data record was associated with a set of attributes for each study participant including race and ethnicity, socioeconomic status, and geographical location – rural, urban, or suburban – in addition to various additional factors relating to the coursework participants engaged in during high school, their college major, and elements pertaining to their interest and engagement in STEM during their schooling.

Due to the time required to complete all transcriptions, data coding was completed in two phases based upon the availability of finalized interview transcripts. The first round of data coding included a preliminary assessment of emerging themes captured throughout the first seven pilot interviews and focus groups, which included 15 individuals. This batch was therefore coded as an exploration sample, which informed the finalized coding schema which would then be utilized for the analysis of the remaining 17 focus groups, including 27 individuals. As a result, the two batches were coded as an exploration sample and a confirmation sample before being combined for analysis.

For the exploration sample, the first step of the coding process involved data reduction. To reduce the data, we conducted an initial read-through of focus groups, while using open coding on the transcripts. We loosely followed a grounded approach and allowed the data codes to develop from the data (Lonkila, 1995). However, unlike strict grounded theory, we did not code the text in a word-by-word fashion, and instead coded sentences or paragraphs with multiple Atlas.ti codes. In this process, we included adequate surrounding text to provide context for the codes used, but also separated portions of text that pertained to distinct themes. As codes in grounded theory are produced through an iterative comparison throughout a qualitative dataset, the coding process continuously evolves. For this reason, after the first round of coding, we read through the data once again with our updated set of codes in order to ensure consistency in our data coding. Multiple members of the study team were involved in the various phases of coding to ensure inter-coder reliability, a technique which best facilitates coding qualitative data obtained from multiple questions for accurate, precise data analysis (Glaser & Strauss, 2017). After coding our exploratory batch of data, we coded the remaining transcripts within our confirmatory sample using the same coding schema. Once the exploratory and confirmatory stages of coding were completed, the data samples were combined for our analysis.

(3) Restricted National Data from the National Center for Education Statistics

In the final year of the *FIRST* Longitudinal Study, we applied for and were approved for access to three restricted databases from the National Center for Education Statistics (NCES): The High School Longitudinal study, the Beginning Postsecondary and Beyond study, and the Bachelorette and Beyond Study.

The **High School Longitudinal Study** collects data for a nationally representative group of over 23,000 graders from 944 schools at baseline in 2009. Students are being followed throughout secondary and postsecondary years. The survey added follow-up waves in 2012, and 2016, and has added high school and postsecondary transcripts to its database. Most of the STEM interest and attitude scales included in the *FIRST* Longitudinal Study used the same measurements included in the High School Longitudinal Study providing the opportunity of direct comparison with this national survey.

The **Beginning Postsecondary and Beyond** study's most relevant cohort began in 2011-2012 with follow-ups 2013-14 and 2016-2017. Students are surveyed beginning in their first year of postsecondary education and followed up at the end of their third and sixth years after entry into postsecondary education. The final dataset for the most recent completed BPS study, BPS 12/17, contains information on approximately 22,500 students. Data collected in this study provide data on enrollment in Engineering and Computer Science courses; college majors, including persistence in major; and degree attained. The data also collect information on employment while in postsecondary and in the year following graduation.

The **Bachelorette and Beyond Study** collects data from a Nationally representative longitudinal study of students who completed the requirements for a bachelor's degree in a given academic year, following graduating seniors 1, 4, and 10 years after completing their bachelor's degree. The latest cohort contains records on over 24,000 individuals. Students in the most recent cohort completed their bachelor's degree in 2015–16 and were followed in 2017 and 2020. Data are now available from the 2020 follow up.

The table below shows the types of variables that match between the *FIRST* longitudinal study and each of the three restricted use datasets.

Exhibit 4: Variables from National Datasets that match the *FIRST* Longitudinal Study

High School Longitudinal Study	Beginning Postsecondary Study	Bachelorette and Beyond Study
Demographics	Demographics	Demographics
Course Taking	College Courses	Job Industry
Honors Courses	College Majoring	Income
STEM scales	Degree Attainment	Job Types
		Job Certifications
		College Majoring
		Clubs/ Competitions/Internships

Propensity Score Matching (PSM)

The national data provide another opportunity for comparing *FIRST* participants to a matched national group through propensity score matching (PSM). PSM aims to equate treatment and control groups with respect to measured baseline covariates to achieve a comparison with reduced selection bias. It is a valuable statistical methodology that mimics Random Control Trials (RCT), creating more equal comparison groups while reducing bias due to confounding factors. PSM can improve the quality of research and broaden the range of research opportunities. PSM is not necessarily a magic bullet for poor-quality data, but rather may allow the researcher to achieve balanced treatment groups similar to a RCT when high-quality observational data are available. PSM may be more appealing than the common approach of including confounders in a regression model because it allows for a more intuitive analysis of a treatment effect between 2 comparable groups.

Typically, there are five steps in utilizing a propensity score matching to equate groups:

1) Collect data
2) Estimate propensity scores
3) Matching
4) Evaluate match quality
5) Evaluate treatment effects (run the analysis)

Collect Data

In order to create an alternative control group for the FIRST longitudinal study, several large nationally representative datasets were acquired for the National Center for Educational Statistics (NCES). These datasets were chosen because they matched data on both demographic factors as well as overlap in outcome measure. Three "unrestricted" datasets were acquired; 1) The High School Longitudinal Study of 2009, 2) The Beginning Postsecondary Students Longitudinal Study (2012-2017), and 3) The Bachelorette and Beyond Study (2016).

Estimate propensity scores

Typically, logistic regression is used to generate propensity scores. The dependent variable used is the treatment grouping (*FIRST* vs control) and the predictors can be any variables that may have an impact on the likelihood of receiving treatment (i.e., being in *FIRST*). With PSM, it is best to use all variables that could possibly have an impact on the outcome. The propensity scores matching procedure for this study used a series of dummy variables for race (white, black Asian, Native American or Alaskan native, and Hawaiian or other pacific islander), geography (urban, suburban, rural), Hispanic ethnicity, English as a second language, and income (above or below poverty level).

Matching

The matching approach used for these analyses is called a "one to one" match in which we use propensity scores to create a control group that has the same sample size as the treatment group. In this methodology, each individual in the treatment group is paired to a similar person with similar demographic and social characteristics.

Evaluate match quality

Once matching is run and both treatment and control groups have been created, it is important to evaluate the quality of the matching procedure. Since the purposes of PSM is to equate groups, if there are still significant differences on demographic factors that would be an area of concern. Comparing the estimates for this study, we find that numbers fairly close, with all measures within one half of a standard deviation.

Evaluate treatment effects (run the analysis)

The final step is running the analysis using the new matched control group on the outcomes of interest. Since the comparison group is matched "one to one" with the *FIRST* group, each grouping contains 822 records for the *FIRST* and matched national group comparison, and 451 records for the *FIRST* Longitudinal study comparison and matched national comparison groups

Limitations of PSM

As with any statistical procedure, it is important to point to the limitations of PSM. Variables known to have significant impacts on the outcome measures, but are not available for matching, can create non-inclusion creates model bias. In our matching procedure, we were able to account for the major factors that we know have an impact on the tested outcomes, such as race/ethnicity, gender, family income, ESL status.

Another limitation of PSM has to do with the matching procedures used. Because of a large pool of potential matches in the national data sets, we used a one-to-one matching procedure "without replacement". This means that once a match is made, that case is no longer available to match. In some instances, the best possible match will be a case that is already matched. As matching "with replacement" creates its own set of potential issues as variance estimation becomes more complex with subjects who are duplicated in the dataset, we opted for the cleaner one-to-one approach.

References

Glaser, B.G., & Strauss, A.L. (2017). *Discovery of grounded theory: Strategies for qualitative research.* Routledge.

Lonkila, M. (1995). Grounded theory as an emerging paradigm for computer-assisted qualitative data analysis. In U. Kelle (Ed.), *Computer-aided qualitative data analysis: Theory, methods, and practice* (pp. 41-51). London: Sage.