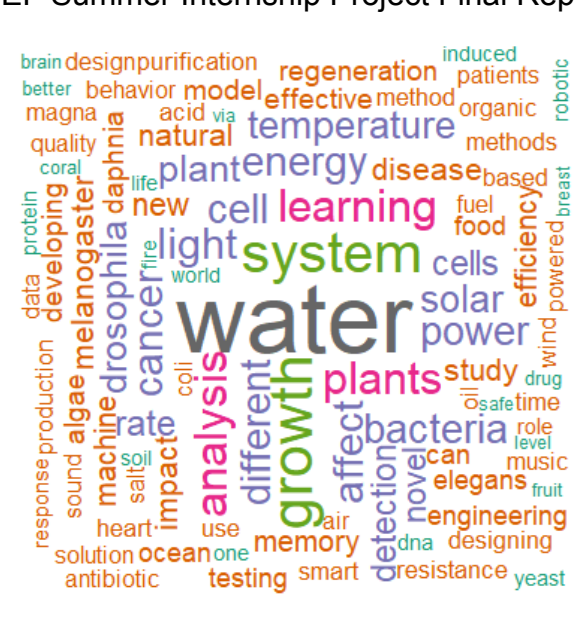
Analysis of Massachusetts Science fair competition participation from 2014-2021 to understand factors driving interest in STEM

Rohan Bandaru
Junior, Lexington High School
2021 MSEF Summer Internship Project Final Report



ABSTRACT

Science fair participation data from over 400 Massachusetts area schools for the years between 2014 and 2021 was analyzed to understand trends in fields of research, student demographics and diversity interest in STEM. Several trends and biases in participation, choice of research topics and performance in the science fair were observed. Some of the key findings were a consistent decrease in participation, increasing gender and ethnic diversity, significant bias in choice of project topics associated with performance and global trends in science and technology. Some of these include an uptick in cancer related research coinciding with the announcement of the Cancer moonshot program in 2016, a significant increase in AI/ML related projects since 2017 applying publicly available deep learning libraries and a smaller and less significant increase in research on infectious diseases with the recent Covid-19 pandemic. A new semantic clustering algorithm was developed and applied to the data revealing interesting patterns in projects and correlation with performance in the science fair. This study resulted in several recommendations to MSEF and participating schools to continue to increase interest in STEM fields and for reducing bias or neglect of fundamental research and certain areas of science and engineering. Also the potential for schools to increase interest in STEM and participation in MSEF without having to increase budgets. Further analysis of this data and continued monitoring of trends in years to come can be a valuable resource for the success of MSEF and participating schools.

1.0 Introduction:

The importance of a strong foundation in STEM has not been more apparent than with the recent economy-crippling COVID-19 pandemic, global climate change, political turmoil fed by extremists exploiting social media and other events impacting us as a civilization.
There is an unprecedented demand for scientists and engineers capable of solving humanity's problems with more and more countries rapidly growing investment in science and technology (Potvin and Hasni 2014). The current race with China to ensure a spot in this tech-centric future (Zahn, 2021) demonstrates the magnitude of this trend.

The Massachusetts Science and Engineering Fair (MSEF) is an organization that aims to encourage an interest in STEM, by allowing students to experience technical and scientific practices by means of original research. Every year, hundreds of students across Massachusetts compete in the science fair bringing several months and sometimes years of innovative research in topics ranging from immunology to electrical engineering, geology, mathematics, chemistry and more. The participants also gain core skills in the scientific method, data analysis and interpretation and presenting results. A skill deeply important in today's society of misinformation, and political extremism. For example, the current anti-vaccine crisis demonstrates the need for better science education (Hornsey, 2018).

A key goal of MSEF's mission is to ensure diversity in the growing STEM community and encourage more participation from the 400 odd schools across the state. The MSEF organization like many others such as WEST (Diperi, 2021) constantly strives to improve and optimize the experience of the participants by implementing more effective programs and broader outreach.

Data from the past 8 years of the science fair competition was analyzed for various trends and patterns. Understanding the trends and biases in the different topics of research, the levels of interest and influence of gender and race on choice of research topics and performance in the competition could help organizations like MSEF to improve their outreach and selection criteria as well as for the participating schools to improve their practices and approaches towards their students (Karampelas, 2020). In this exploratory exercise, I sliced and diced the data, looking at all kinds of factors such as the topic of choice, gender, race and various school level characteristics for over 2000 science fair entries by roughly 3500 students from 400 high schools over the past 8 years. In addition to MSEF, the role of the schools in encouraging participation in STEM and in the competition is crucial (Gonzalez, 2012). Analysis of the schools baseline attributes and how they are faring in comparison to their peers in not only participation but also the performance of the projects in the competition can be a valuable guide to the school administrations in investing more resources or taking corrective actions.

1.1 Motivation and personal statement

This analysis was done in part as a summer internship program that I did with MSEF in the summer of 2021 along with some other students. I had since also taken an AP statistics course and I revisited this data to apply my newly acquired statistical analysis skills. I found this data set fascinating given I had myself participated in the MSEF challenge and wanted to explore and understand as much as I could of this data. Encouraging STEM participation in the student community has been a long passion of mine and I was engaged in a similar activity through another initiative www.codingsafari.org over the past 3 years. The analysis of this data set I felt would be valuable to help me further that cause as well as also provide MSEF and other organizations insights into factors that could be impacting interest in STEM.

2.0 Methods:

2.1 Data Description:

The raw datasets covered 8 years of MSEF science fairs, from 2014-2021. The data was received in two tables for each year of the fair from 2014-2021. In the first table, each row represented a participant in the fair, with the columns containing their project title, grade, gender, ethnicity, school, school region, and school zip code. The second table for each year had each row as a project, with the columns including project category, and project placement. These two tables were merged for each year to create one single dataset containing every participant/project that participated in MSEF 2014-2021. There were about 2000 unique research projects from various fields of sciences, engineering and math. The data also included a broad categorization of the projects into engineering or chemistry or environmental sciences etc. This classification was not always accurate and was probably based on what the participating students had self selected as the category of their project. Data was provided at the level of each individual participant, with data on gender, grade and race. The overall performance of the project (placement) was also available. Across all 8 years, there were 2541 individual participants with 9 attributes on the research project. The identity of all individuals was anonymized in the dataset provided by MSEF.

Data on the participating schools was extracted from the statewide rankings of the schools from https://www.schooldigger.com/go/MA/schoolrank.aspx?level=3. This site provided school enrollment and student demographic summaries and overall school ratings/rankings for 351 public schools. This data was merged with the subset of schools that were also in the MSEF project participation data set.

Year	Students	Projects	Schools	Boys	Girls
2014	393	306	78	186	186
2015	387	304	77	171	192
2016	386	314	67	167	187
2017	378	304	78	151	203
2018	375	304	86	165	178
2019	364	293	77	131	199
2020	330	265	87	131	197
2021	315	256	65	No Data	No Data

Table 2.1.1: Summary of counts for the data set used in this analysis. A more detailed summary of the statistics for all the different variables in the data set is available in the supplementary material.

2.2 Data Processing and Normalization:

Errors in school names, region of participation and other missing data points were manually fixed and cleaned up. Missing entries were interpolated or inferred from other entries in the dataset. Additional labels were then added to the data provided by MSEF, including project topics, relevant terms describing the project available from the title.

2.2.1 Consolidation of continuation projects:

Several projects were continued and re-submitted year after year by the teams but with slightly updated titles. These projects were identified by simple rules comparing the nearest match on the topic of the project, the school, gender and ethnicity composition of the participating team which all added to the probability of the project being the same one as entered in previous years. This was used to generate unique project ids. For each project that had multiple records additional attributes such as team size, gender and ethnicity composition of the teams were calculated. Also the projects were reclassified into engineering, biological and chemistry, math and computer science and social sciences fields based purely on the project titles. This was done manually. For projects that could not be discerned into a specific field of science, I used the original field provided by the participants when submitting the entry. I applied a text clustering approach (described in 2.4.1) which helped identify duplicate project entries and removed or tagged them appropriately for all further analyses.

2.2.2: Consolidated performance scores:

Performance scores at the level of the school were calculated by pooling all projects submitted from a school in any given year and generating a weighted score based on the placement of the projects. Projects that placed 1st, 2nd and 3rd contributed, 1, ½ and ⅓ to the score respectively. Projects with honorable mention contributed 1/8th to the score and just having an entry (no awards) contributed 1/20 for each project. These values were arbitrarily chosen. All scores were normalized for the number of entries from the school.

2.2.3 Additional derived variables and data harmonization

One interesting observation in the data was the choice of research subject. There seemed to be some tendency to choose topics that were more applied or had a high impact on society. To test this additional classification and categorization of the projects was done. A score from 1 -10, with 1 reflecting fundamental or basic research to 10 being research deemed as of immediate applied value or impact on society was given to each project. This determination was based solely on the project title description so it is likely that there were errors. At the time of analysis, to reduce any bias in this score, the values were binned into 3 categories (1-4) as pure/basic/fundamental research, 5-7 as intermediate, 8-10 as high impact applied research. Projects were also categorized in multiple ways, one was the field of research or the skills needed to conduct the project, another was the field of application of the research. For example, there were projects that applied machine learning and computational methods but the application was in the medical field. Also all the projects were broadly classified into three types, those that conducted experimental analysis exploring existing data or generating new data, projects that developed an innovative method or approach on a research problem and projects that actually developed a device or mechanism or application.
Another dimension to this data was the ethnicity of the students which was used to identify diversity in the student participation in MSEF. In order to test if the students participating in MSEF reflected the underlying diversity or make-up in the student community in the schools, a Shanon Wiener diversity index was used for each school across all the 8 years of data. The same index was calculated for each school where data was available.

2.3 Limitations and Assumptions:

This exercise was aimed as an exploratory analysis to identify trends and patterns in participation in the science fair and to ultimately influence the organizers and policy decisions to increase interest and participation in STEM fields. Given that, the data set used for this analysis has several limitations.

1 . It only represents the fraction of students and projects that made it past the first round of selections at the individual school level. And trends seen in this data are likely to be influenced by that selection bias.

2. The data spans 8 years while the school attributes data was only available for 2018 and 2019, but the assumption is that general school/student demographics have not changed significantly over the 8 years.

3. Project classification data was not very descriptive, and was determined by the students entering the project. Many projects spanned several different subject areas of science. For example application of AI/ML methods in medicine spanned skills and expertise in both Computer science and Biology/medical fields. Even after some clean up and adding additional ontologies to the data, there is likely to be some error and bias in the classification of the projects.

4. Statistical testing of the trends and any biases observed was not corrected for the multiple testing and so the observations are likely to be impacted by false discoveries. Given this was an exploratory analysis, the assumption is that any trends or patterns seen in this analysis should be further tested and explored with additional or independent data and analysis.

5. Another limitation of this data is the geographical representation. While not only limited to the state of Massachusetts, even within Massachusetts, the student representation is highly biased to the Boston Metro and surrounding areas. This can be a challenge in interpreting and developing any practical policy changes beyond Massachusetts. However, one advantage to this constraint is that the analysis is less influenced by broader variability and confounding due to differences in geography, socio-economic, political and cultural differences. Still, the findings from this analysis can be very valuable and useful to other educational agencies and organizations beyond Massachusetts for policy decisions and programs to enhance STEM skills in their communities. They will just have to keep in mind these specific limitations underlying the data that was used for this analysis. Wherever possible, I have tried to validate the observations with broader global sources.

2.4 Analysis methods and tools:

This data was analyzed at multiple levels: the individual student level, the project/team level, and finally the school-wide level. All data was processed and analyzed in MS Excel, Python 3.0 and R Studio 6.0 software.

At the individual project level, analysis was conducted for testing associations of the project categories, participant gender, ethnicity etc. on the performance and choice of topic etc. A second analysis at the schoolwide level was conducted to study the association between school attributes such as ranking of the school, number of students, student to teacher ratios, school funding etc. on the participation and performance of the school in the competitions. To explore the trends in the project topics, a word cloud of titles was created after removing basic English stop words as well as commonly occurring but not relevant words like "effect" and "using". The word cloud was created with the R packages tm and SnowballC. Specific trends in the research topics and association with broader global trends were tested by identifying projects in specific topics such as Cancer, AI/ML and projects related to Covid/infectious diseases etc.

To test for associations between project topics, and performance and other categories, Chi-Square tests for association were used at significance level of 0.05.
Testing for bias in performance or participation based on certain factors such as gender etc. was done using a Z-test of proportionality also at a significance of $P<0.05$.
Simple correlation plots for longer term trends and associations between ranks and performance score or student body size vs participation were used to look for general associations.

Analysis of the relationship and representation of the ethnic diversity in MSEF was done by focusing on key racial groups and combining the other groups as "other" to ensure sufficient numbers were available for analysis. The ethnic groups in MSEF data were defined as White, Black (African American), Hispanic and Asian (which included South Asian and Asian) and others. The School level data obtained had similarly White, Black, Hispanic, Asian, American Indian, Pacific Islander and two or more races. For each school a diversity metric was calculated using the Shanon Wiener diversity index given by Summation of $pi * log(pi)$ where p_i is the proportion of the i'th ethnicity. This analysis was done by pooling all the 8 years of data.

2.4.1: Clustering Analysis:

Unsupervised clustering analysis of the projects and the schools was done using two different approaches. Projects were clustered using a categorical data clustering method that I developed following some published approaches (Koch, 2020 and Allahyari et. al. 2017). This approach calculated all pairwise semantic distances of the projects based on key words in the project title and the area of research. In order to calculate the semantic distance that was meaningful, all terms used in the project titles were synonymized using a public dictionary (From: https://www.gutenberg.org/ebooks/51155). Then additional terms describing the ontology of the words in the title were added. For example, the project title:
"Effect of Coffee on fruit fly depression", would become "effect caffeine, organic, chemical, natural, d.melanogaster, insect, organism, depression, addiction, disease, neuroscience, behavior". This ensured that this project would cluster with other projects looking at the effect of chemicals on organisms at the highest level, and within that cluster closely with organisms being insects or even deeper to specifically fruit flies, or the cluster could be driven by the area of disease, neuroscience to specifically depression. To remove this ambiguity in clustering, all the terms including the ontological terms I added were weighted manually. Terms that were more generic (verbs and generic nouns) got a lower weight = 1, while terms that were more specific (mostly nouns, objects) got higher weights, with very specific nouns like the names of organisms or companies or products got the highest weight = 4.

In addition, the weights were further normalized by the frequency of their occurrence across all the projects. The choice of weights and formulae for calculating the distances were optimized through several iterations on a small subset of the data. I plan to use more semantic clustering approaches available publicly that depend on machine learning and natural language processing capabilities in the future. (Chablani, 2019 and Kulmanov et. al., 2021). This exercise was to take a more controlled approach to cluster the projects.

For calculating the distances, I took the reciprocal of the squared weighted sum of shared terms, normalized to a 0-1 scale and then added to this the dissimilarity measure that accounted for the total number of terms between pairwise comparisons. That is the number of terms that were not shared. The distance matrix generated was then clustered using a neighbor joining tree algorithm. Other attributes such as the schools, the team size, gender and ethnicity composition of the team were mapped on to these clusters. The text processing and distance calculations were performed using a series of Python scripts and the tree generation on the distance matrix was done using R.

3.0 Results:

This data set presented some very surprising results. Interestingly, the highest frequency of words in the project titles seems to be "water" as seen in the word cloud at the beginning of this article. It is not apparent if this was something to do with a common theme of research in the last decade, perhaps it reflects the recent water shortage and crisis faced on the west coast (Becker, 2021) . There were a large number of projects relating to themes on environment and saving wetlands, desalination technologies, clean water etc. This requires additional investigation. But other highly frequent terms were around Machine Learning, Cancer, Energy, Bacteria. These were further explored in the results below.

3.1 Trends in participation and choice of topics over time:

We see a consistent trend of decreasing participation in the Science fair over the 8 years. The Covid-19 pandemic has definitely impacted participation as shown in the table below, but there appears to be something more than Covid-19 itself that might be driving this decline.

3.1.1 Participant Gender and Performance in Fair:

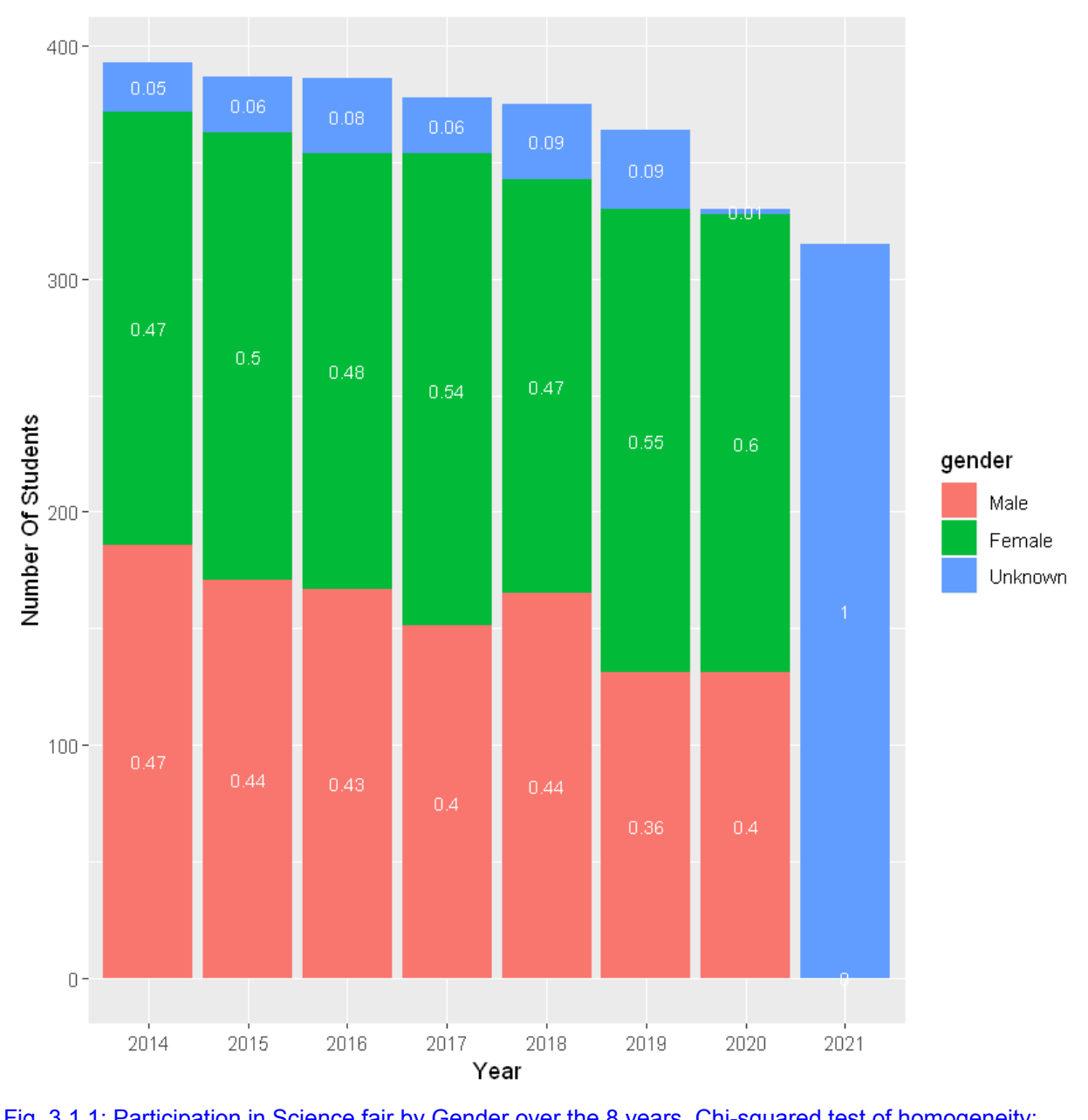


Fig. 3.1.1: Participation in Science fair by Gender over the 8 years. Chi-squared test of homogeneity: p-value = 0.02607

It appears the drop in participation is mostly due to the male students (Fig 3.1). There has been a consistently higher number of female participants across 7 years except 2014 (where it was evenly split). Also worth noting is the decrease in male participants, whereas female participants have remained relatively the same. This trend was not statistically significant, but it does show a slight decrease in overall number of male students participating in the science fair.

3.1.2 Participant Ethnicity and Performance in Fair:

There is a noticeable increase in the participation by South Asians, Asians, Hispanics and African Americans over the years, which points to a trend to increasing diversity in the participation (Fig. 3.1.2)

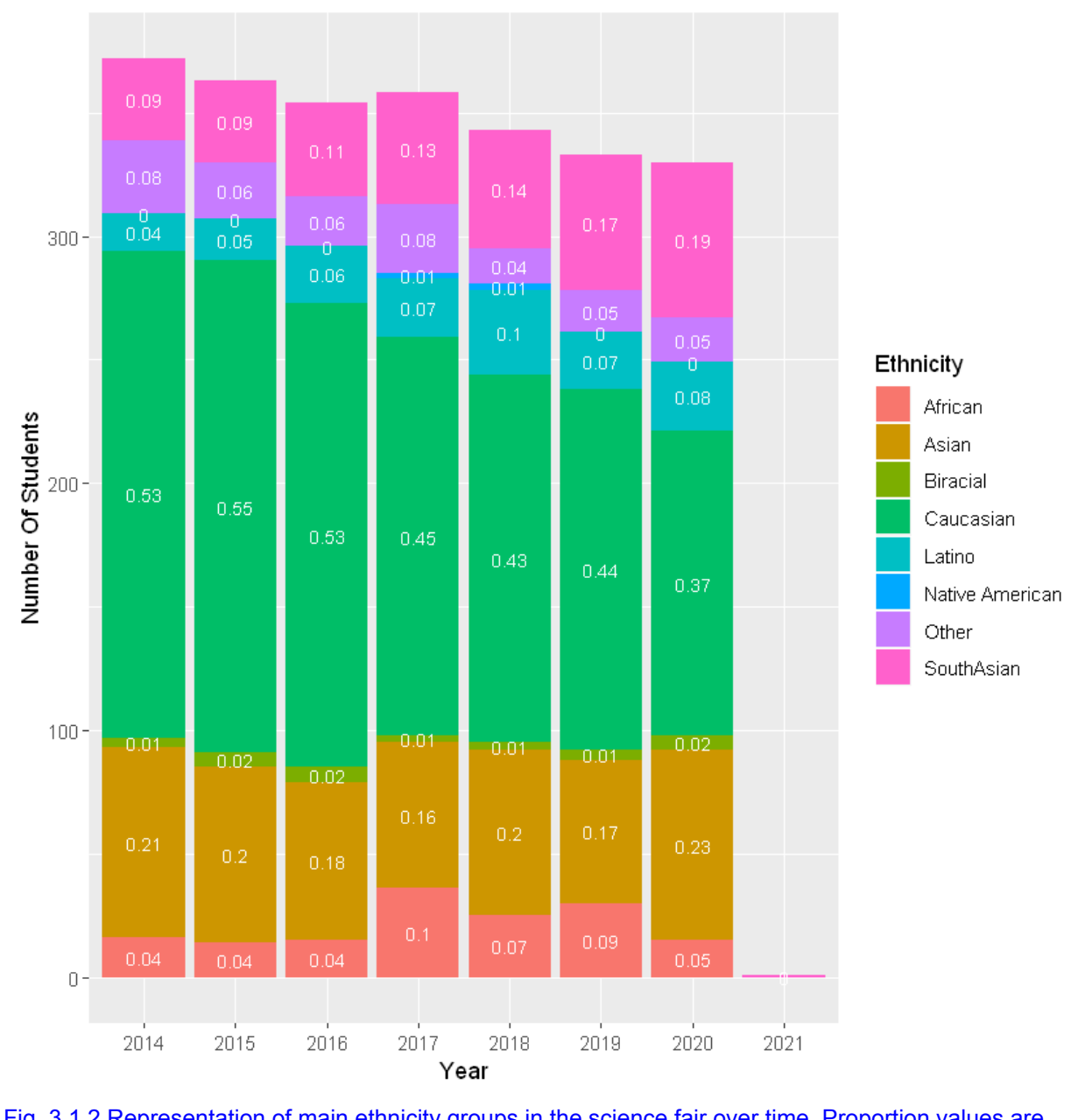


Fig. 3.1.2 Representation of main ethnicity groups in the science fair over time. Proportion values are shown within the bars.

3.1.3 Trends in project topics:

The choice of topics for the research project seem to be influenced by certain world events and this is clearly seen with the sharp increase in the number of research projects on infectious diseases, specifically looking at Covid-19 related research. Similarly there was a spike in projects on Cancer topics in 2017. A substantial increase in application of AI/ML approaches was seen in the last 4 years. All these trends are likely associated with current world events. Chi-squared test for homogeneity: p-value = 1.766e-08

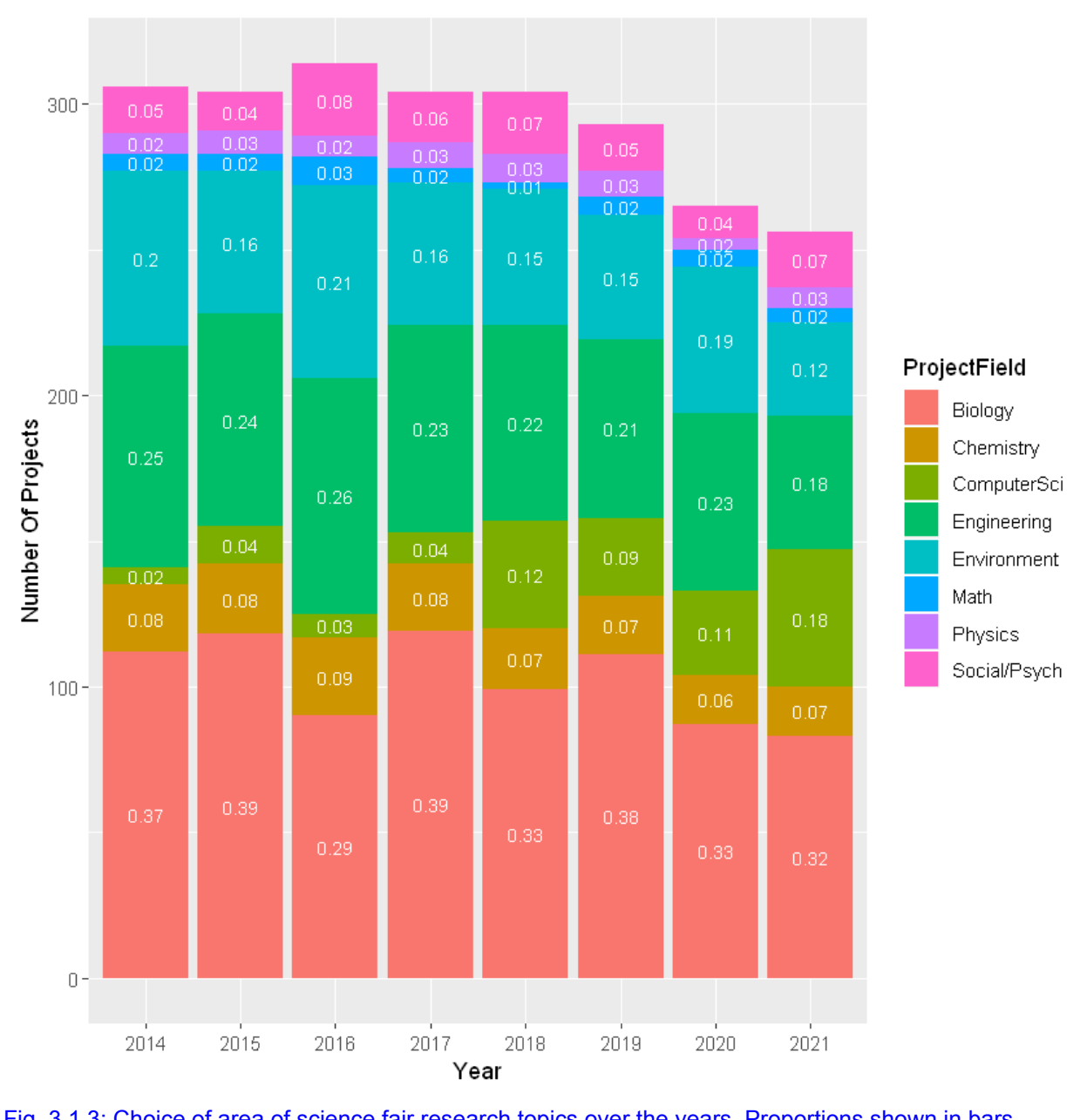


Fig. 3.1.3: Choice of area of science fair research topics over the years. Proportions shown in bars.

Overall, there seems to be no apparent trend in the field of science, except for a significant increasing trend for projects related to computer science topics which is likely contributed mostly due to a sharp increase in AI/ML related projects since 2017.
Chi-squared test for homogeneity: p-value = 2.837e-14

3.1.4 Projects involving Infectious Disease:

There is a spike in the year 2021, with both the proportion of projects and awards increasing for projects related to infectious diseases and is likely attributable to Covid-19 Pandemic.
Chi-Squared test: p-value = 0.4723

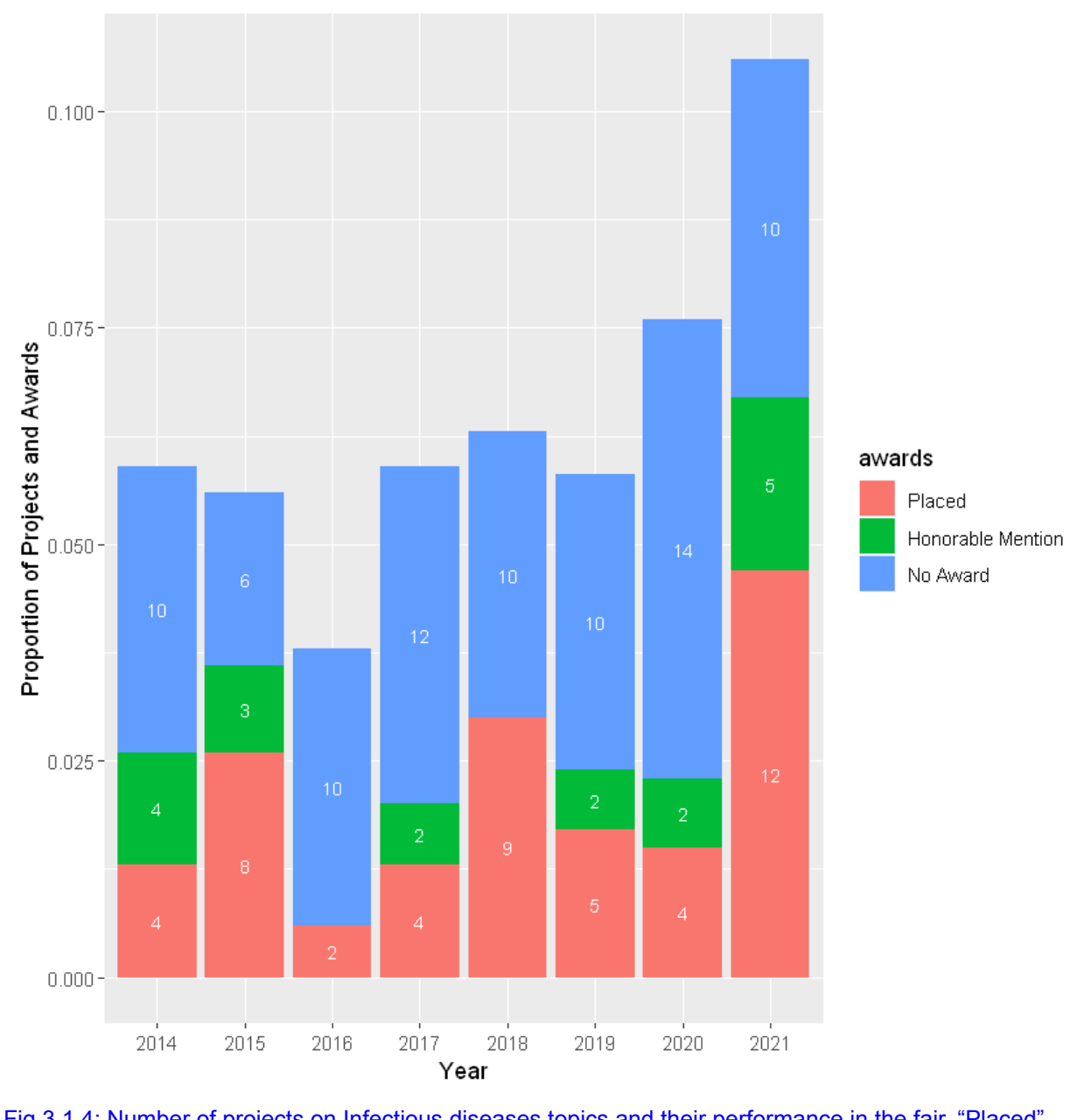


Fig 3.1.4: Number of projects on Infectious diseases topics and their performance in the fair. "Placed" includes 1st, 2nd and 3rd placing projects. Frequencies shown in bars.

3.1.5 Projects involving artificial intelligence (AI) and Machine Learning (ML):

There is a significant upwards trend in AI/ML projects since 2018, and the frequency of projects is increasing year after year. Chi-Squared Test: p-value = 5.37e-13

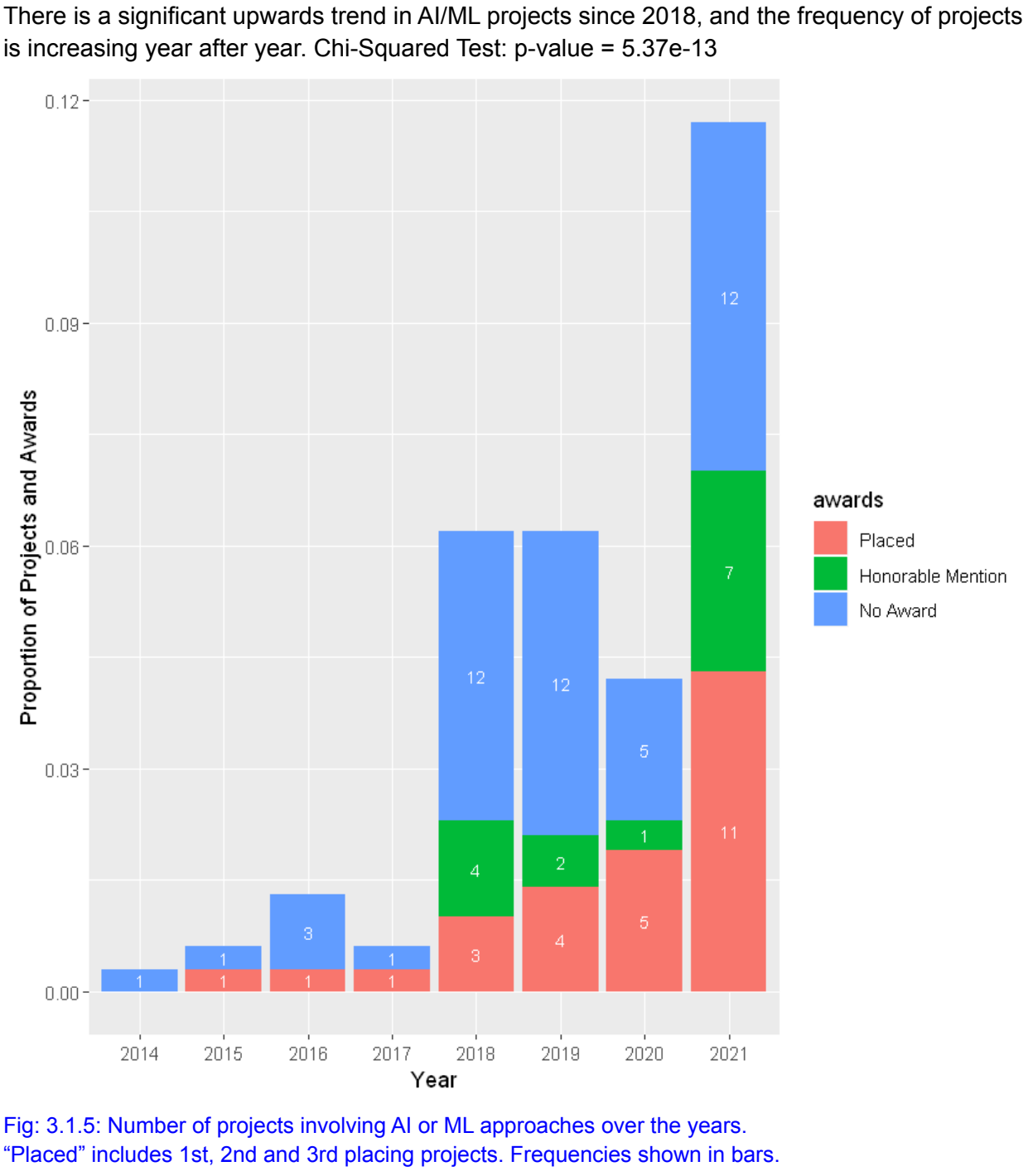


Fig: 3.1.5: Number of projects involving AI or ML approaches over the years.
"Placed" includes 1st, 2nd and 3rd placing projects. Frequencies shown in bars.

3.1.6 Cancer-Related Projects:

Due to the small sample size of Cancer projects, it's tough to establish any significant trends. However there is a large spike of projects in 2017. Chi-Squared Test: p-value = 0.02571

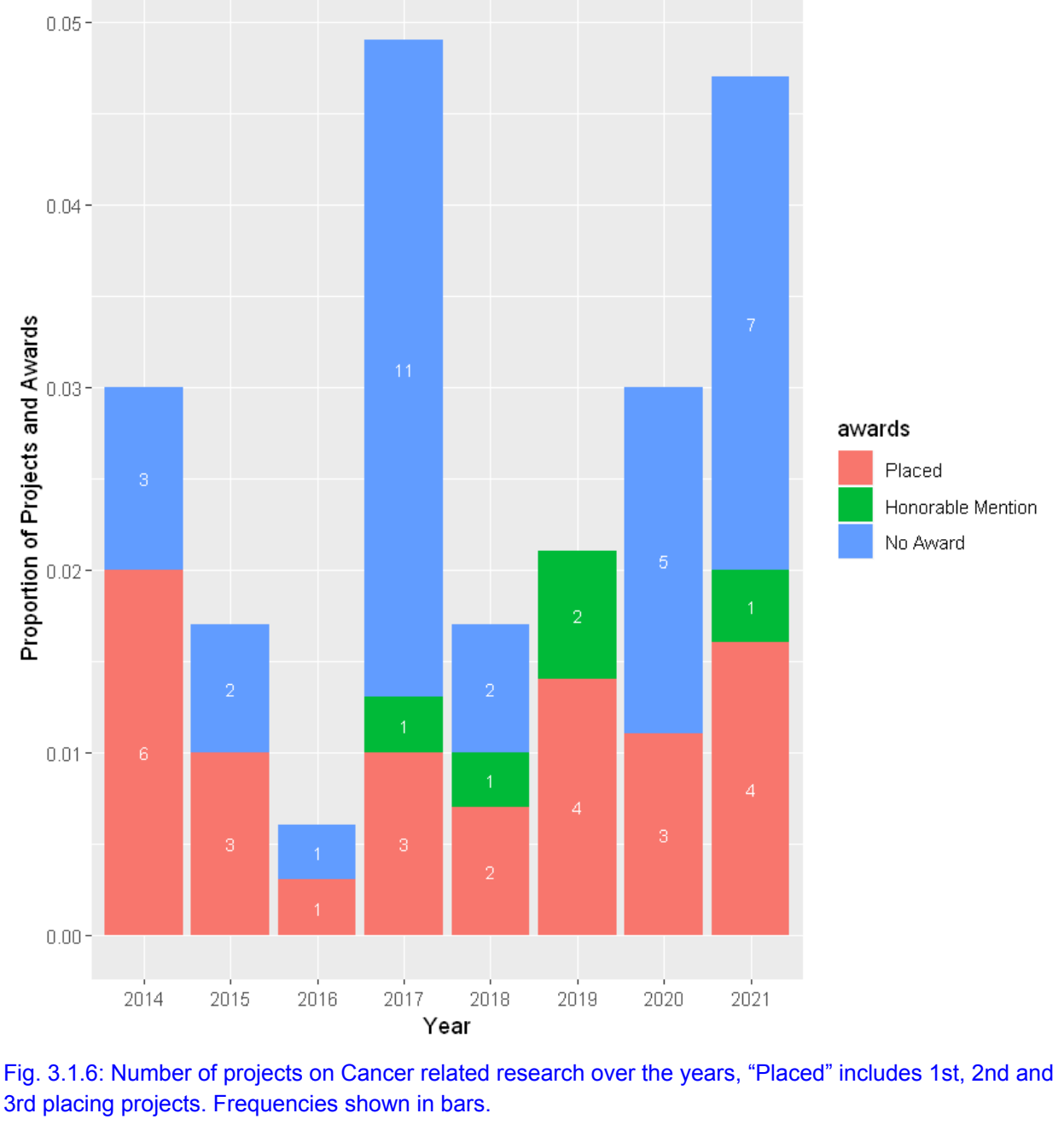


Fig. 3.1.6: Number of projects on Cancer related research over the years, "Placed" includes 1st, 2nd and 3rd placing projects. Frequencies shown in bars.

I also looked at other trends over time such as the number of participating schools, regional distribution (whether there was an expansion of geographic regions) over time etc. But these analyses did not show clear trends and were also biased by a lot of missing data. There was also no significant trend or differences over time in the age (grade) of the students participating in science fair.

3.2 Biases in participation:

3.2.1 Field of research by gender:

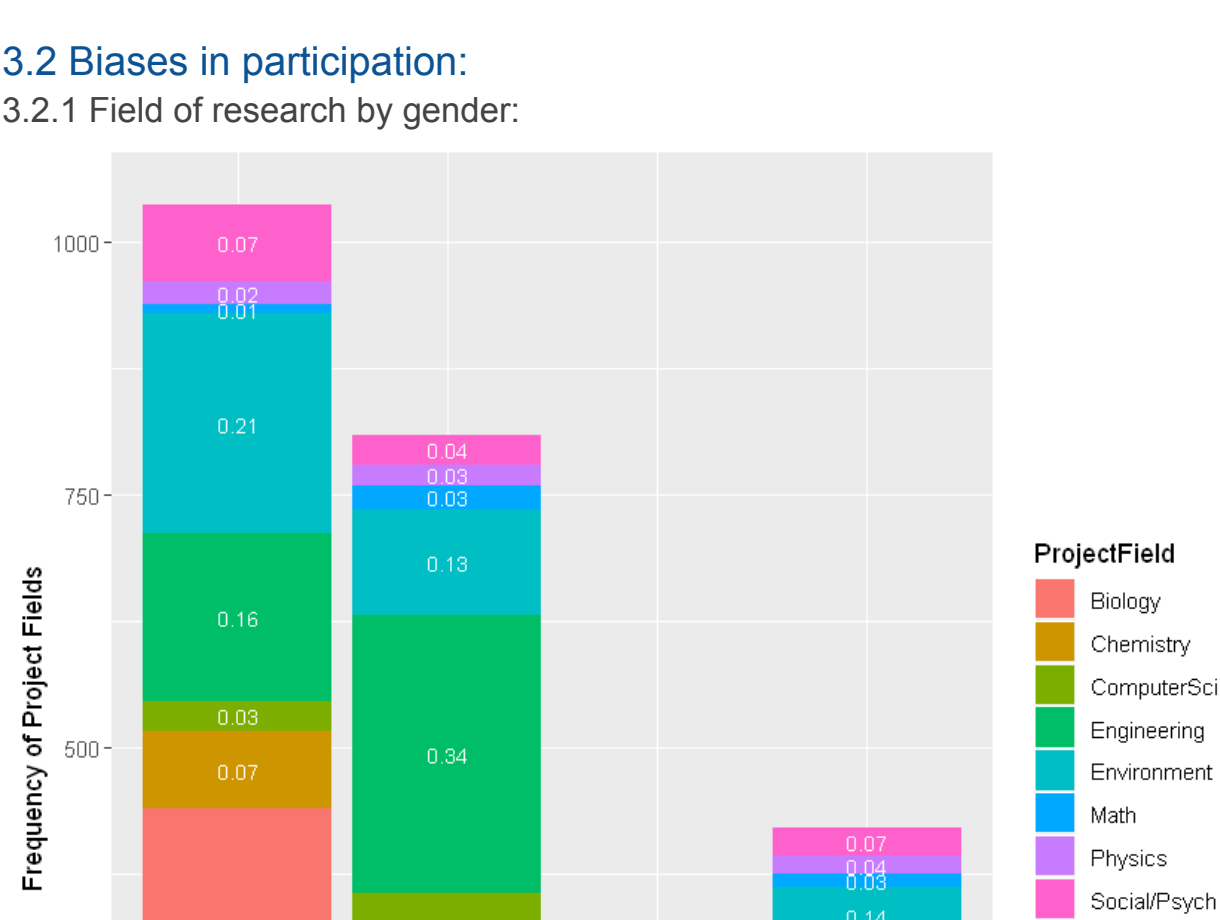


Fig. 3.2.1: Association of field of science fair research and gender of the researching student.

Preference of project topics were significantly (p < 0.001) influenced by gender of the participants, with almost twice as many biology projects done by female participants than male participants, while the opposite is the case with engineering projects (16% vs 34%) and almost 4 times as many computer science related projects were done by males as compared to female participants. These trends were consistent across all 8 years. With a significant proportion of projects having no gender data, these trends could, in reality, be more or less pronounced.

3.2.2 Grade of the participant and field of research:

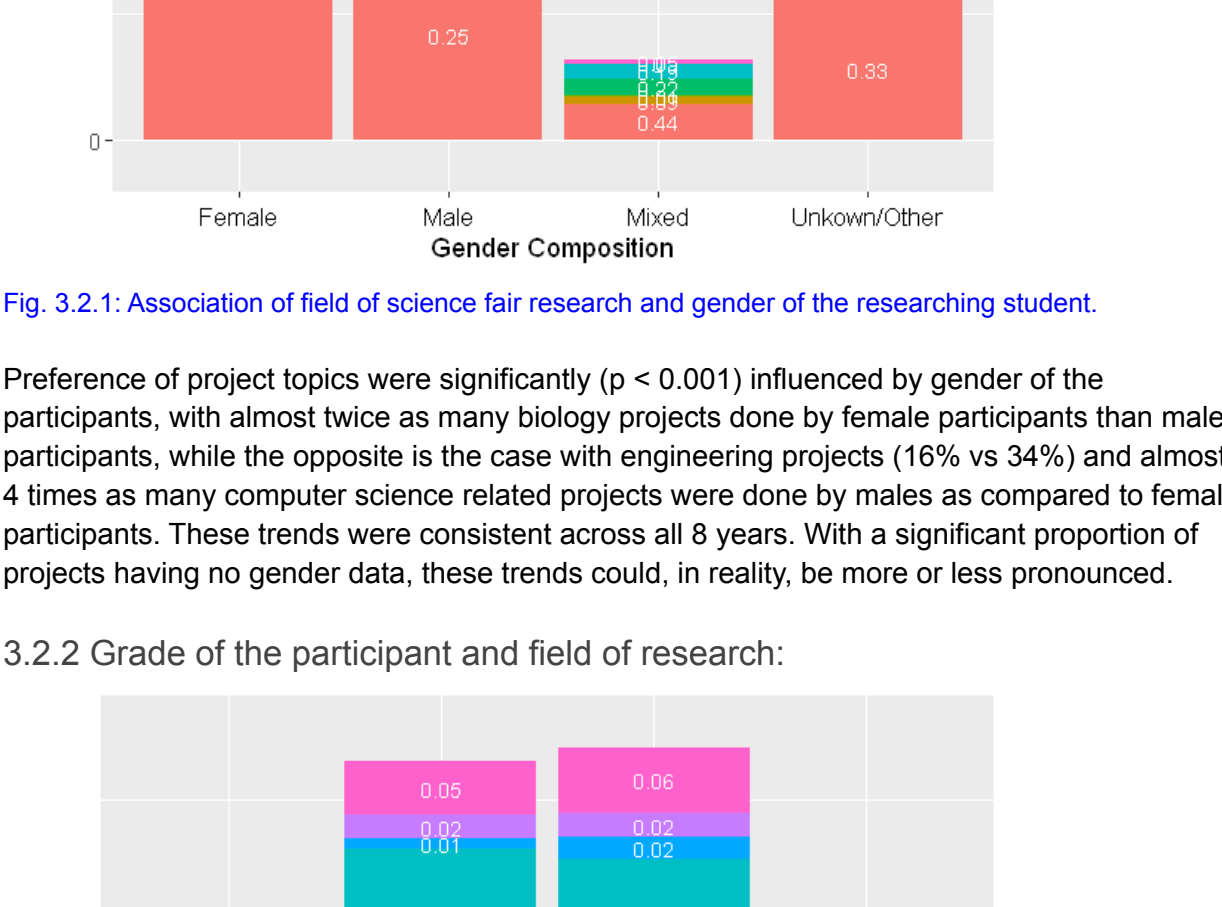


Fig. 3.2.2: Association of field of science fair research and grades of the researching students.

3.3 Factors affecting performance:

As someone who has participated in MSEF and in discussions with other participants, I realized that students often select projects not only based on their interests, but choose topics that could be favorable for awarding. I wanted to test whether this has any merit, and if certain fields or formats perform better than others.

3.3.1 Impact of participant grade on award:

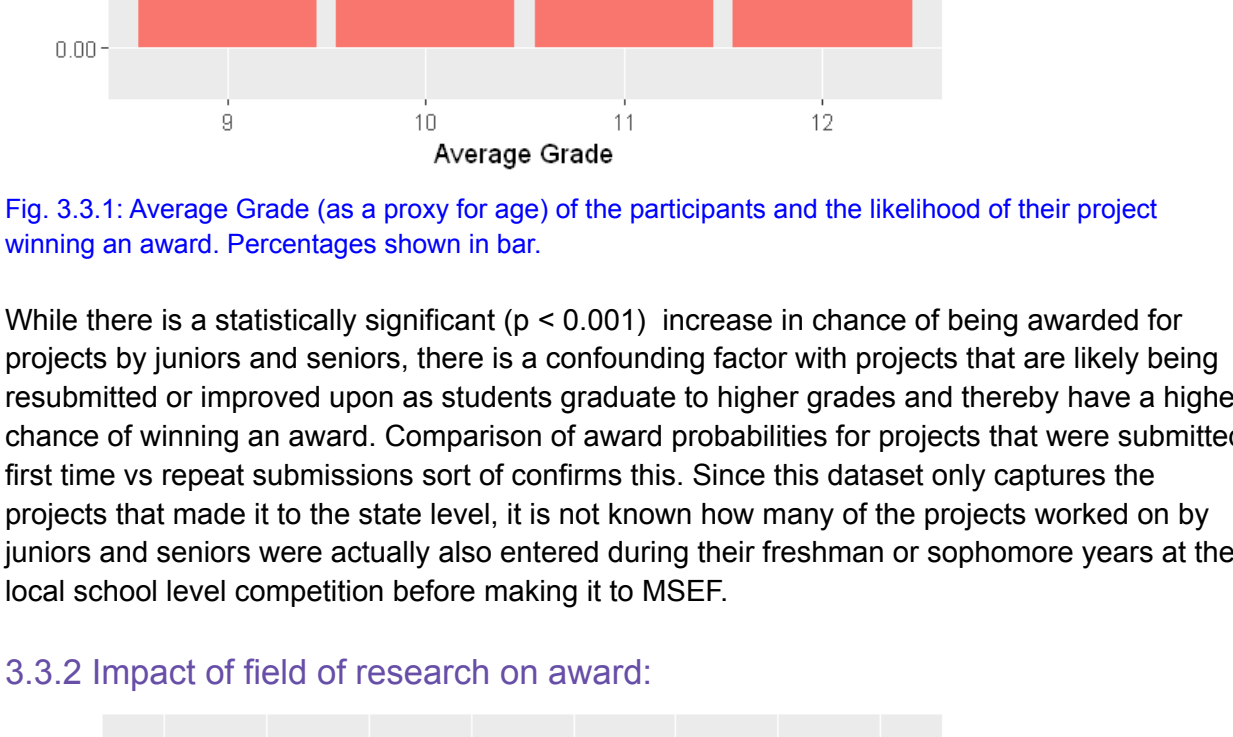


Fig. 3.3.1: Average Grade (as a proxy for age) of the participants and the likelihood of their project winning an award. Percentages shown in bar.

While there is a statistically significant (p < 0.001) increase in chance of being awarded for projects by juniors and seniors, there is a confounding factor with projects that are likely being resubmitted or improved upon as students graduate to higher grades and thereby have a higher chance of winning an award. Comparison of award probabilities for projects that were submitted first time vs repeat submissions sort of confirms this. Since this dataset only captures the projects that made it to the state level, it is not known how many of the projects worked on by juniors and seniors were actually also entered during their freshman or sophomore years at their local school level competition before making it to MSEF.

3.3.2 Impact of field of research on award:

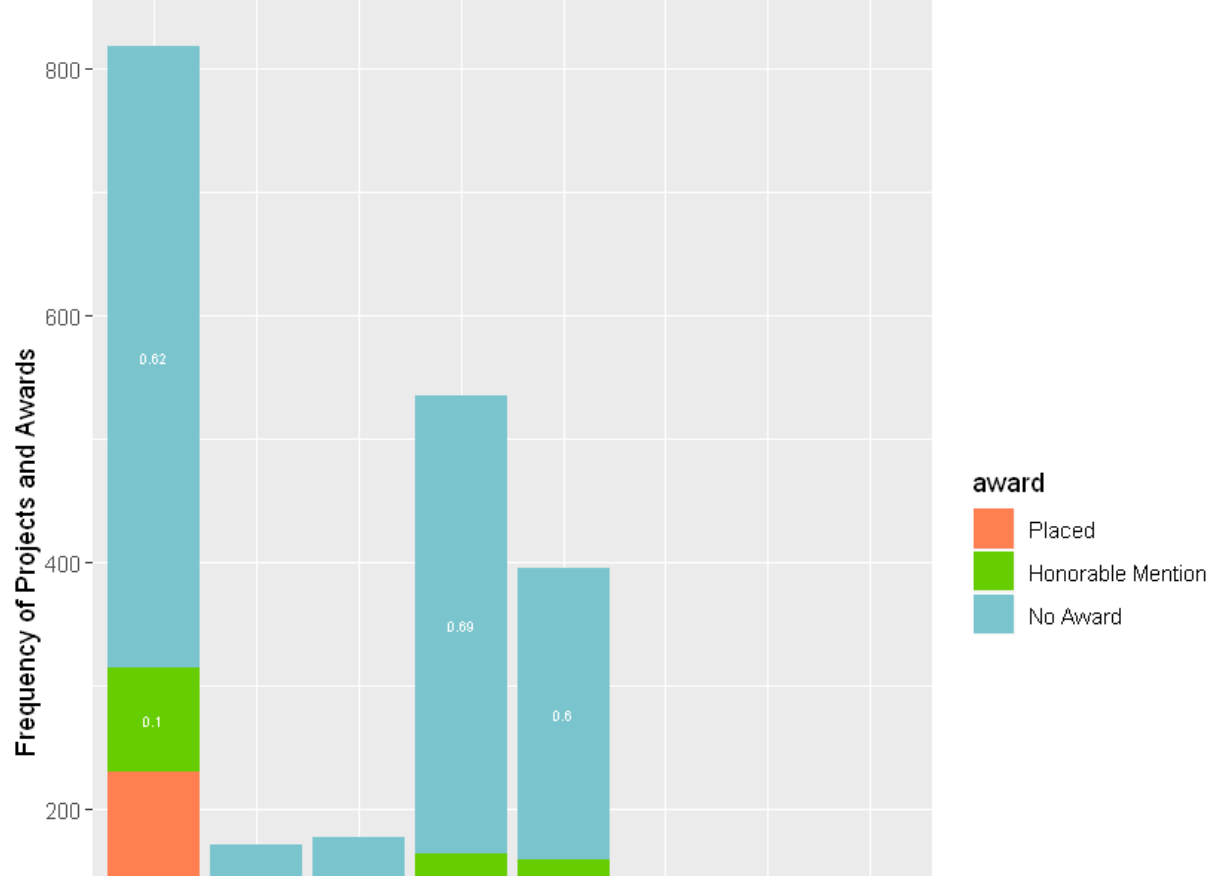


Fig. 3.3.2: Probability of winning an award in the different fields of research. Proportions show in bars.

Projects in Computer Science seem to have the highest probability of being awarded a placement or honorable mention (49%). Projects relating to Biology, Environment or Behavioral/Social topics have close to 40% chance of being awarded. Projects in pure Chemistry, Math or Physics typically perform poorly. Furthermore, choice of specific fields or topics such as Cancer or AI/ML topics seems to definitely increase your probability of winning an award (p < 0.001)

3.3.3 Impact of certain key project areas on award:

I looked at key project topics related to "Cancer", "AI/ML", "Water conservation", "sustainability" and "infectious diseases" and some of these seem to show a clear advantage over the other more fundamental research problems in being selected by the judges for awards.

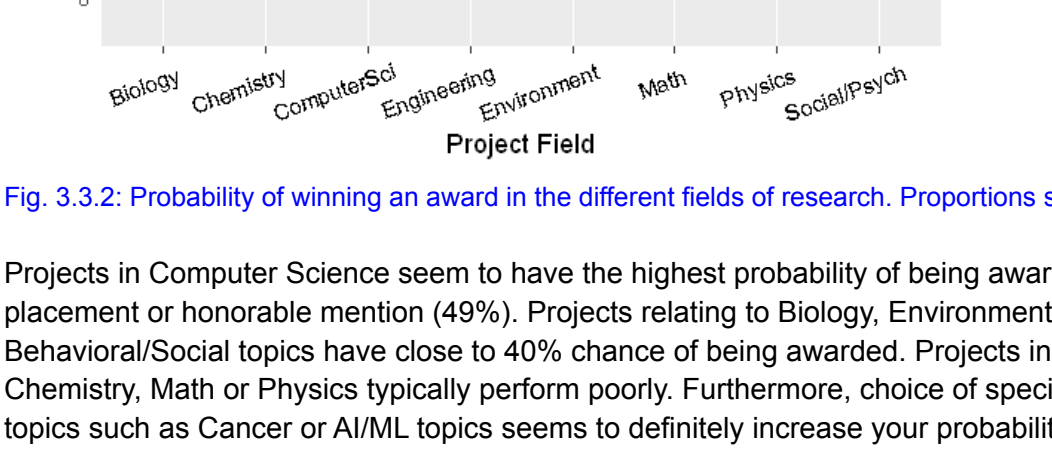


Fig. 3.3.3a: Proportion of projects winning an award "Placed" = (1st, 2nd or 3rd place) for projects related to Cancer (bar on right) and all other projects (bar on left)
Fig. 3.3.3b: Proportion of projects winning an award related to AI/ML (bar on right) vs all other projects (bar on the left).

There is a significant increase in the chance of getting an award if a project relates to Cancer, or AI and Machine Learning.

3.3.4 Applied vs Fundamental research topics:

Note that, when the projects were characterized on this scale on the applied value of the research, it was done purely based on the project title description. So there is likely to have been some error in this classification. However, the idea was to generally test if students were favoring fundamental research or more applied research topics. This data was also used to test if projects perceived as having greater applied value are more likely to be selected for awarding.

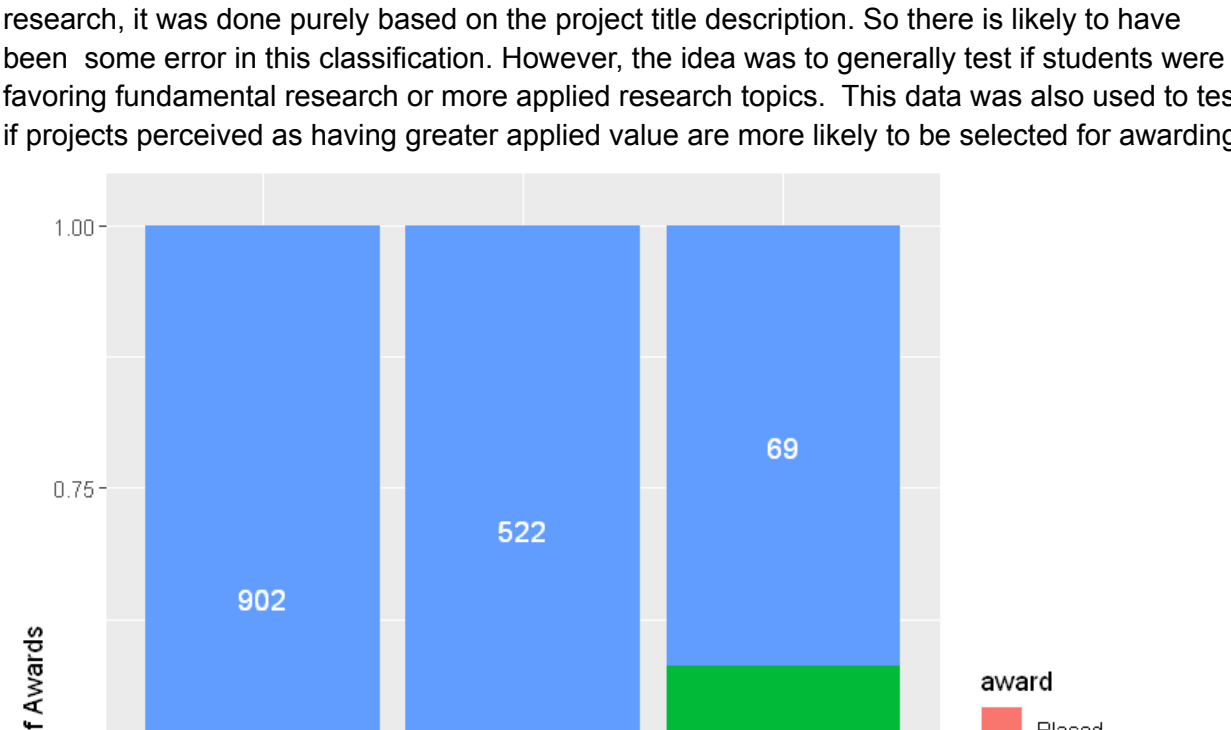


Fig. 3.3.4: Likelihood of winning an award (1st, 2nd or 3rd place) or HM = honorable mention for projects on basic/fundamental research/ low applied value (1), medium (2) and high applied value (3). p-value = 6.634e-15. Frequencies shown in bars.

3.3.5 Project team size on probability of placing:

Most of the science fair entries were by individual participants, but there were quite a few projects with 2 or more team members working on projects. Are projects that had larger teams and thus likely to have done more work, more likely to win an award?

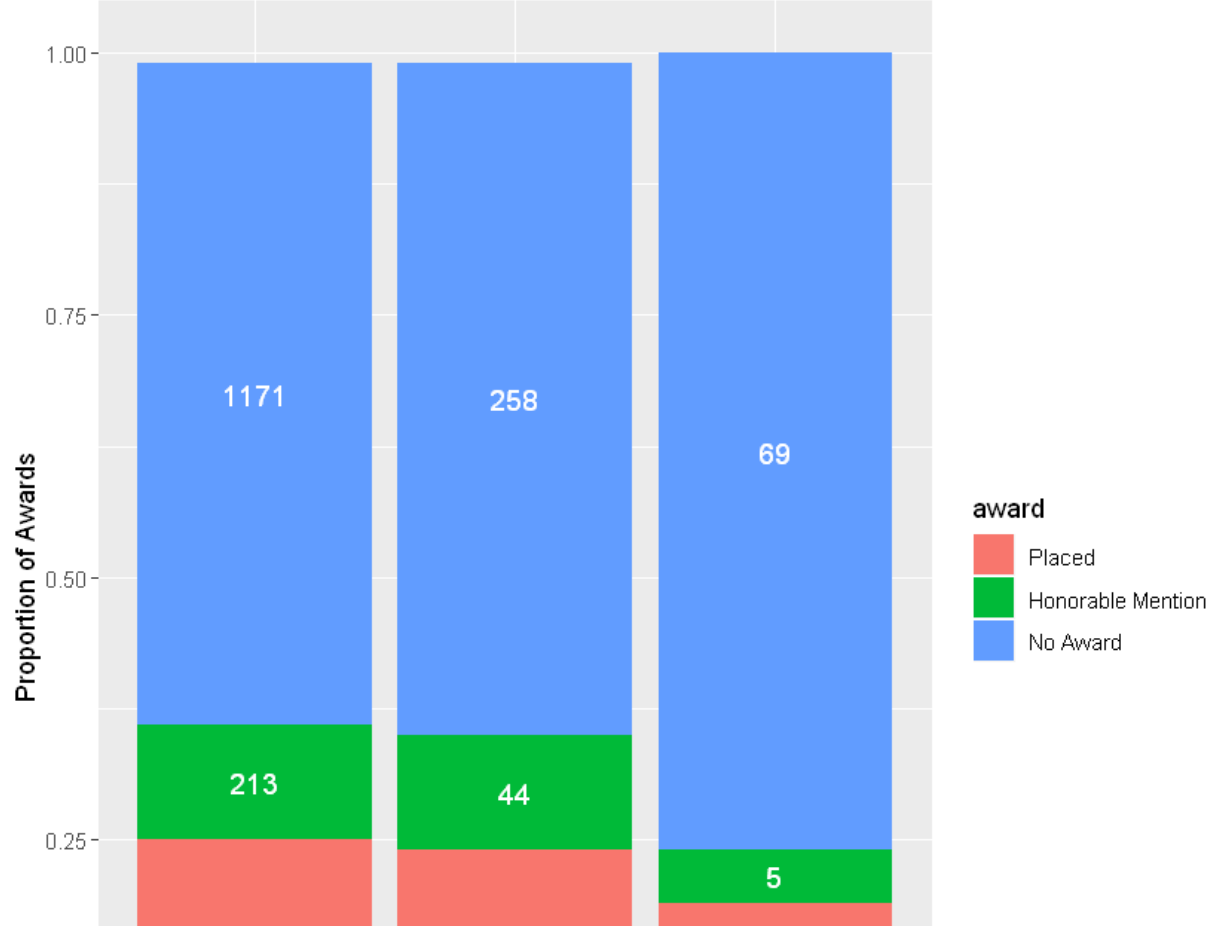


Fig. 3.3.5: Effect of team size on likelihood of winning, 1 = single person project, 2 = 2 members worked on the project and 3 = 3 or more?. p-value = 0.157

Having more members working on a project does not seem to have any advantage in terms of increasing the chance of being awarded, if anything the opposite seems to be likely.

3.4 School level Participation and Performance:

I also looked at participation and performance in the science fair aggregated at the school level. There were a total of 378 different schools in the state of Massachusetts represented in the science fair with only around 70-80 schools entering the competition in any given year between 2014 and 2021. Thus not all schools had an entry every year. A few schools dominated the participation with multiple entries every year. As shown below there were schools (Lexington High School) that had a total of over 150 entries over the past 8 years as seen below in fig. 3.4.1.

3.4.1 Number of Projects from Schools:

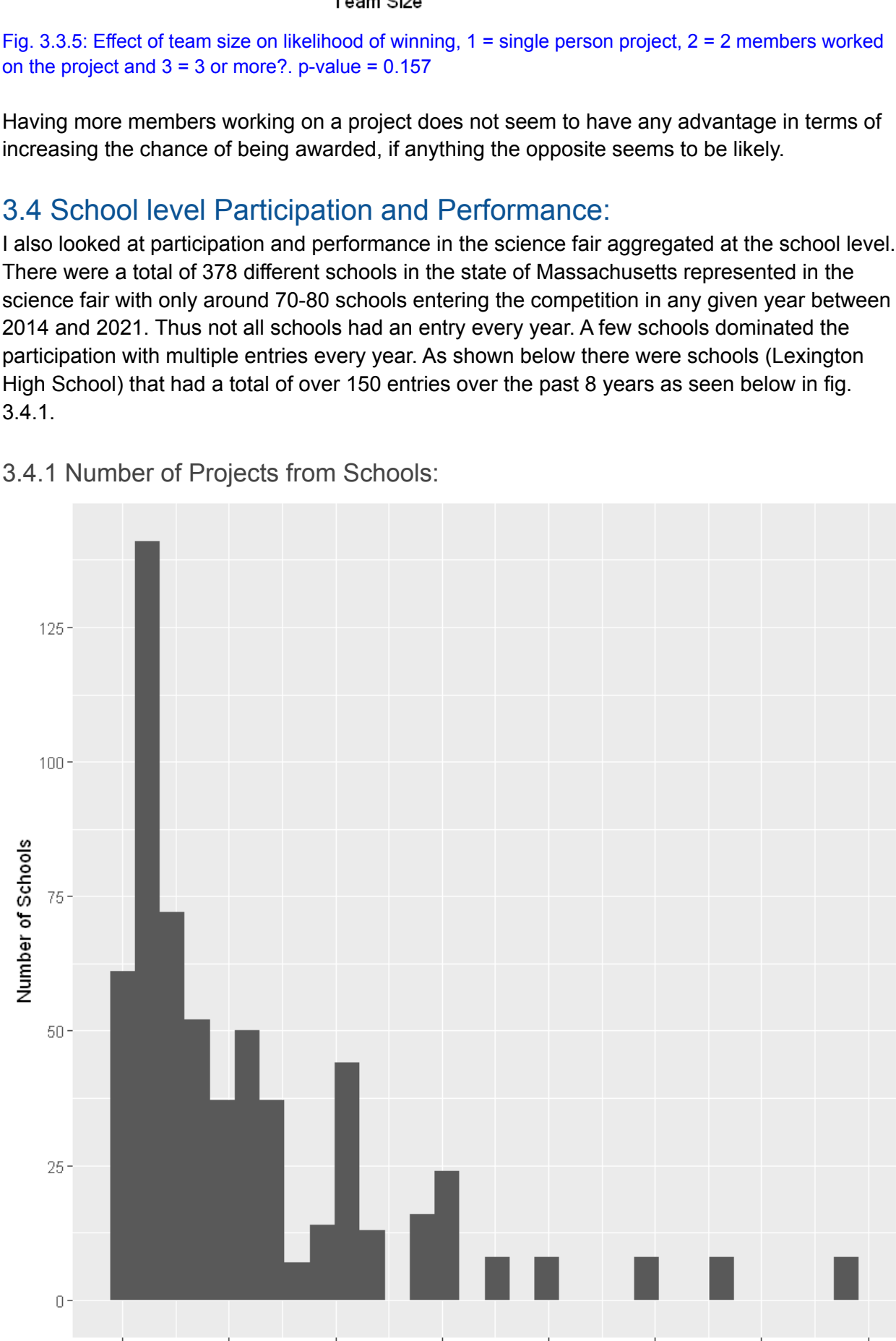


Fig. 3.4.1: Distribution of the number of entries from schools. Most of the schools had only 1 or two entries in the past 8 years and a few schools account for a large participation each year.

3.4.2 Participation of Schools by School Size:

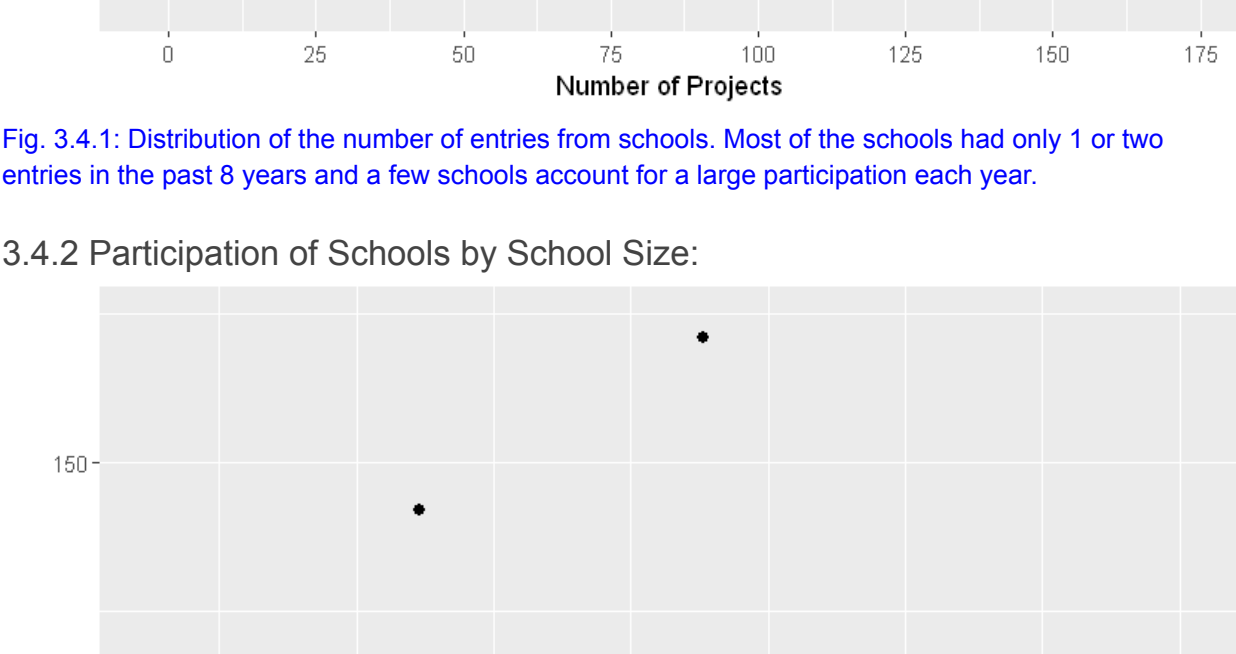


Fig. 3.4.2: Size of school (# of students enrolled in 2019) vs the total number of projects submitted by the school between 2014 and 2021. Note: Many schools did not have a submission in 2019, so it was decided to pool the submission data over the 8 years for the Y axis.

The number of projects submitted to MSEF from a school does not seem to be related to the number of students in the school or to other factors such as the funding the school gets or the overall performance ranking of the school based on academic performance of the students. Although there is a slight positive relationship and the slope is statistically significant (P < 0.001).

3.4.3 School Performance Vs Rank:

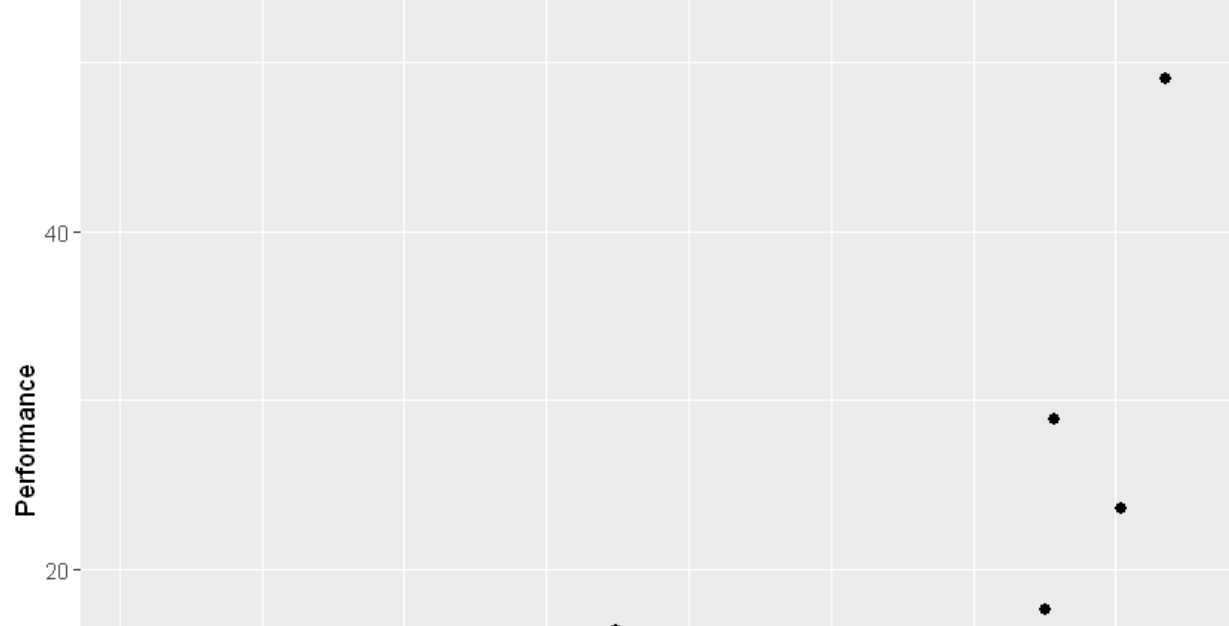


Fig. 3.4.3: School rank (as listed on SchoolDigger) vs the performance score (across all years) of the school. Performance score is a weighted average of the awards for all the projects that the school, so schools that win more awards have higher scores. Outlier in the top-right corner is Lexington High School.

There is a slight positive correlation (P < 0.001) showing that the academic ranking of the school has some small component to performance in MSEF.

3.4.4 School Funding vs Participation:

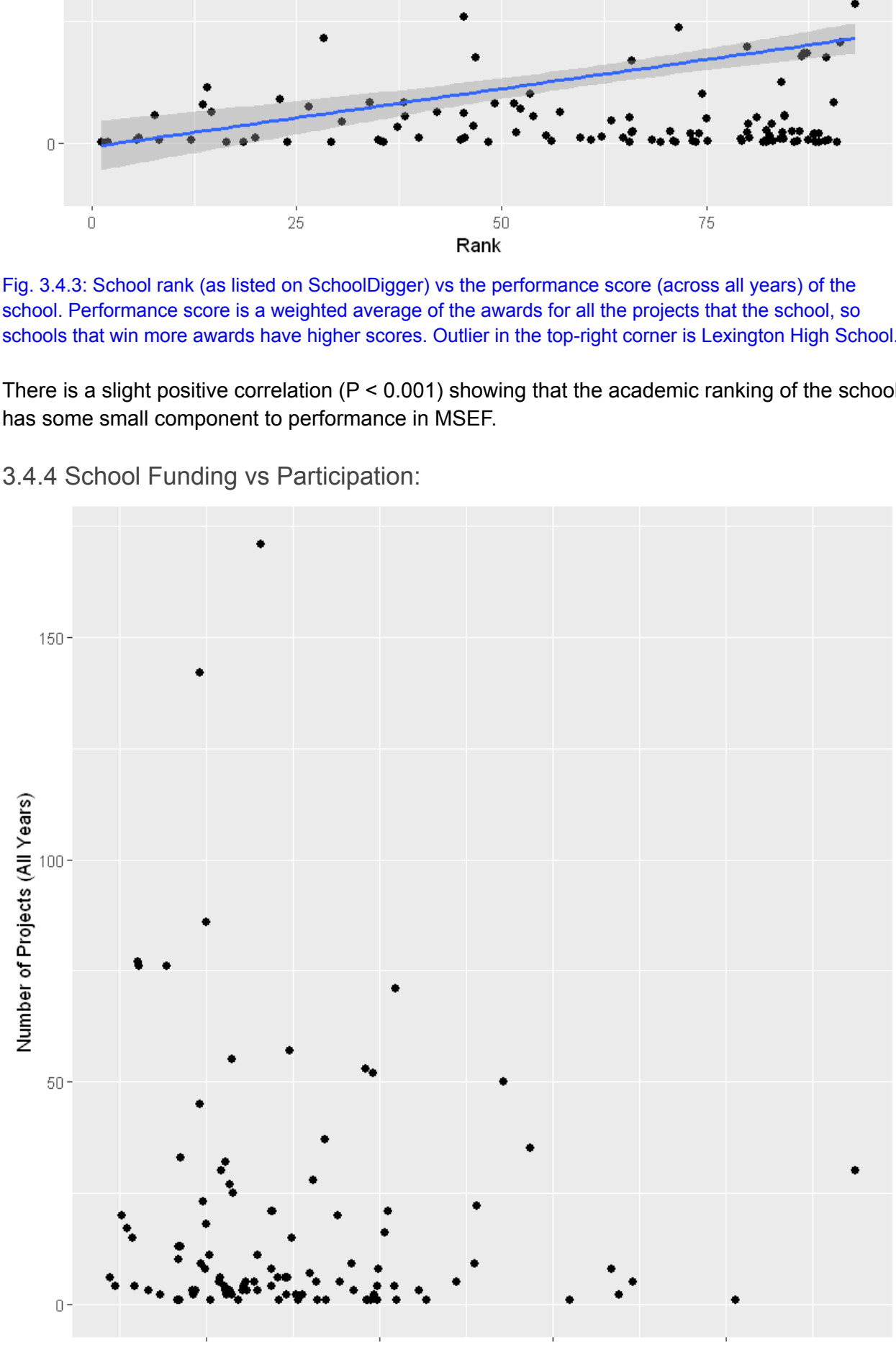


Fig. 3.4.4: School per-student funding (source: SchoolDigger.com) vs the total number of projects entered by the school across all years.

Ignoring the schools for which there was no spending data (Spending = 0), there is no apparent correlation (P > 0.05).

	Schools Ranked by Performance	Ranked by Participation
1	Massachusetts Academy of Math & Science	Lexington High School
2	Lexington High School	Westfield High School
3	Wachusett Regional High School	Massachusetts Academy of Math & Science
4	Falmouth Academy	Falmouth Academy
5	Hopkinton High School	Hopkinton High School
6	Bancroft School	Wachusett Regional High School
7	St. Mark's School	Shrewsbury High School
8	Shrewsbury High School	Taunton High School
9	Westfield High School	Bishop Feehan High School
10	Bishop Feehan High School	Somerville High School

Fig. 3.4.5: Top ten schools by overall performance and participation across all 8 years.

3.5 Clustering analysis of the projects by semantic similarity

So far we have looked at various attributes of the projects and the participants individually mostly through univariate analysis. The goal of the clustering exercise was to explore and understand if there were more broader patterns in the data that were driving both the participation and the performance. By taking an unsupervised clustering approach that clustered the projects solely based on certain attributes, it is possible to analyze the resulting clusters for enrichment in participation or performance. This approach takes away the rigid categorization that was applied to the projects and instead allows for a multivariate unbiased and independent determination of project types. This same approach was also applied to the schools, clustering them based on their participation and performance in MSEF and other attributes. The clustering algorithm developed for this exercise seems to be performing very well and in general brings together projects similar in the nature of their research topics. It also helps to quickly identify duplicate records or projects that were submitted repeatedly over the years with minor modifications to the title. But more importantly the clusters generated seem to have some correlation with probability of being placed or awarded. Some of the clusters are also enriched for specific student attributes such as gender and grade of the participants. Fig. 3.5.1 shows a small example of this clustering algorithm applied to all projects on physics related topics. The distance metric was not rescaled from the global clustering of all projects and so no importance should be given to the actual distances in the dendrogram.

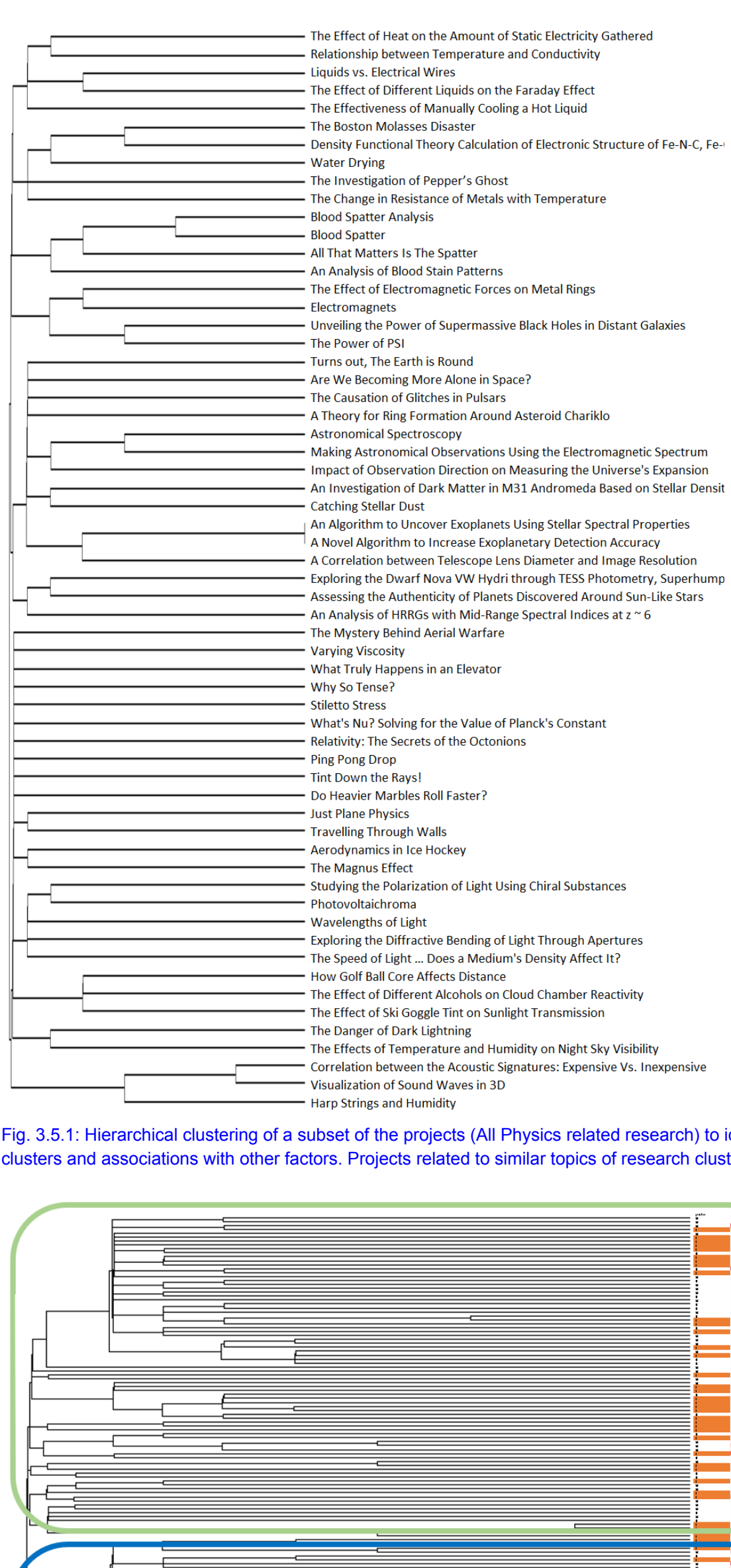


Fig. 3.5.1: Hierarchical clustering of a subset of the projects (All Physics related research) to identify sub clusters and associations with other factors. Projects related to similar topics of research cluster together.

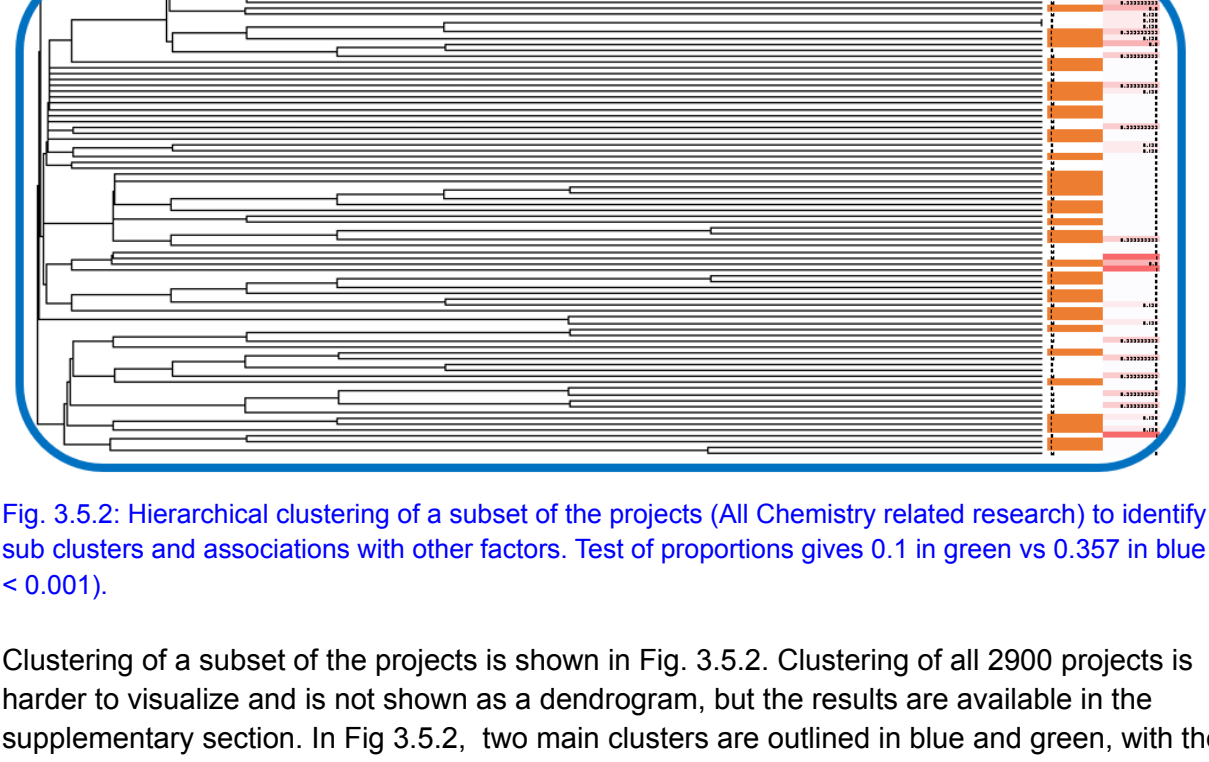


Fig. 3.5.2: Hierarchical clustering of a subset of the projects (All Chemistry related research) to identify sub clusters and associations with other factors. Test of proportions gives 0.1 in green vs 0.357 in blue (p < 0.001).

Clustering of a subset of the projects is shown in Fig. 3.5.2. Clustering of all 2900 projects is harder to visualize and is not shown as a dendrogram, but the results are available in the supplementary section. In Fig 3.5.2, two main clusters are outlined in blue and green, with the projects in the blue cluster winning more awards, about 30% more (red markings) while the green cluster has less awards. Similar observations were seen in the larger clustering results. While other factors like Gender and grade also showed some heterogeneity, in the clustering above of all Chemistry related projects, there was no statistically significant disparity in gender preference (Projects with female participation highlighted in orange) to sub clusters.

3.6 Summary of all analysis results:

I looked at several aspects of the science fair projects, analyzing trends over the 8 years for participation, diversity of the student participants, choice of research topics and also looked at biases between various categories of students and project areas. Table 3.6 below summarizes some of these comparisons that were conducted and the key take away or inference from each of the tests. I tried to use appropriate statistical tests to test the hypothesis in each case where possible and other situations just looked for trends or patterns in the data visually. Section # in table 3.6, corresponds to the Results section of this report where more detailed results of the comparison were described. All the results are discussed in the discussions section.

Section #	Comparison/ Analysis	Test and Results	Inference
3.1.1	Gender bias in participation over time	Chi-Sq test: **p < 0.05*** n = 2928	Decrease in participation over the years driven by drop in male students.
3.1.2	Bias in Ethnicity representation over the years in Science Fair	Chi-Sq test: **p < 0.001***** n = 2454	Increasing diversity over the years, significant reduction in Caucasian participation over the years.
3.1.3	Trends in subject area of research	Chi-Sq test: **p < 0.01**** n = 2346	Significant increase in Computer Science projects over the years.
3.1.4	Trends in infectious disease related topics	Chi-Sq test: p < 0.4 n = 2346	Noticeable increase in the frequency of infectious disease projects in the years 2020-2021, and increased award.
3.1.5	Trends in projects involving AI/ML	Chi-Sq test: **p < 0.001***** n = 2346	Significant increase in the frequency of projects, and the awarding of projects relating to AI/Machine Learning.
3.1.6	Trends in Cancer related topics	Chi-Sq test: **p < 0.001***** n = 2346	Large spike in Cancer-related projects in 2017, with an increase in the number of Cancer-related projects in recent years.
3.2.1	Bias in choice of subject area for research by Gender of participant	Chi-Sq test: **p < 0.001***** n = 2346	Significant gender bias in project research topic. Biology projects are female-dominated, whereas Computer Science projects are male dominated.
3.2.2	Bias in choice of subject area for research by Grade of participant	Chi-Sq test: **p < 0.001***** n = 2346	Increase in the proportion of biology projects from Freshmen to Seniors. Computer Science projects peak in 11th grade.
3.3.1	Grade of participant and likelihood of winning an award	Chi-Sq test: **p < 0.001***** n = 2186	Significant increase in the likelihood of winning an award for Juniors and Seniors. The probability of winning peaks in Junior year.
3.3.2	Subject area of research and likelihood of winning an award	Chi-Sq test: **p < 0.001***** n = 2346	Computer Science and Biology projects have the highest chance of winning. Math and Physics do poorly.
3.3.3	Choice of specific topics related to cancer, AI/ML and likelihood of winning an award	Cancer: Chi-Sq: **p < 0.01**** n = 2346 AI/ML: Chi-Sq test: p < 0.2 n = 2346	Significant increase in the chance of winning an award when a project involves Cancer, but this trend is insignificant for AI/ML
3.3.4	Applied vs fundamental research and likelihood of winning awards	Chi-Sq test: **p < 0.001***** n = 2339	Significant increase in the chance of winning an award when the applied value goes up.
3.3.5	Size of the participating team and likelihood of winning an award	Chi-Sq test: p < 0.2 n = 2346	Slight decrease in the chance of winning as the group size increases.
3.4.1	Number of projects from schools	N/A	The vast majority of schools send few projects (<30) with a few schools sending upwards of 130 projects.
3.4.2	Number of students enrolled in the school vs the number of MSEF participants from the school	T-test of Slope **p < 0.001***** n = 614	Very weak upwards correlation between the total students enrolled and the number of MSEF participants.
3.4.3	Overall performance of the school in MSEF and the academic ranking of the school	Spearman's Correlation: **p < 0.001***** n = 614	Very weak upwards correlation between the awards won by the school vs the academic ranking.
3.4.4	Level of participation in MSEF and the funding/resources of the school	Slope t-test: p < 0.5 n = 614	No visible correlation between the level of participation and per-student funding of the school.

Table 3.6: Summary of the various comparisons and analyses conducted on the MSEF data set. Stars represent significance. *** is highly significant and * is marginally significant. P-values reported are unadjusted.

4.0 Discussion:

A key goal of MSEF is to encourage participation in STEM, while also increasing diversity in these fields. This study aimed to analyze how these goals have been accomplished, and identify potential areas for improvement. It provides insights into the participation and choice of research topics by the student community, and how this affects their performance in MSEF. These insights can be very valuable to students and schools looking to get involved in the fair. Overall, MSEF participation has become more diverse, both with student demographics and the projects themselves. However, there are still some trends that need attention.

4.1 Trends in participation and research topics:

One trend that raises concern is the steady decrease in overall participation in the science fair. In recent years, interest in STEM in general seems to be decreasing (Potvin, P. 2014), so this trend could be reflected in MSEF participation as well. The declining trend could also be due MSEF perhaps being more selective, with fewer students from the regional fairs making it to the state level. Additional data from the regional competitions may help to resolve this. The decrease was definitely exacerbated by the Covid-19 pandemic, however the decline in participation is apparent from earlier years as well. The decline in participation is not due to few schools participating. Rather, it seems to be partly driven by a drop in male students entering the fair, starting out with an equal split in 2014 and trending towards more female students participating. It is not clear if this is specific to Massachusetts or a wider or global phenomenon. National-level analysis of data revealed a significantly lower rate of pursuing STEM fields among female students (Kong et. al. 2020). It is nice to see the reverse trend in the MSEF participation data, implying that whatever the schools and MSEF are doing to encourage female participation in STEM research might be working. Similarly, student ethnicity has shown positive trends, with increasing participation from minorities and an overall increase in diversity over the years.

Looking at participation by grade, the almost 2x increase from 9th to 10th grade makes sense, since many students don't find out about MSEF well into their 9th grade and aren't prepared to participate. However, the 50% drop from 11th to 12th grade is surprising. I expect a large part of this decrease is due to students being busy with college admissions. Most seniors are busy completing their applications, not to mention that many participants are motivated by the prospect of having this on their applications, which would be over by the time MSEF happens. However, for the 12th graders that do participate, they are the most likely to place in the fair. This expected upwards trend in performance from 9th to 12th grade confirms that awards increase with experience and skill set.

Over the years, certain project topics have become more or less common in the fair. For example, Computer Science projects have seen a steady increase throughout all 8 years, while general Engineering projects have been decreasing. In particular, specific topics, such as AI and Machine Learning, have seen drastic growth in recent years, as it becomes more relevant.

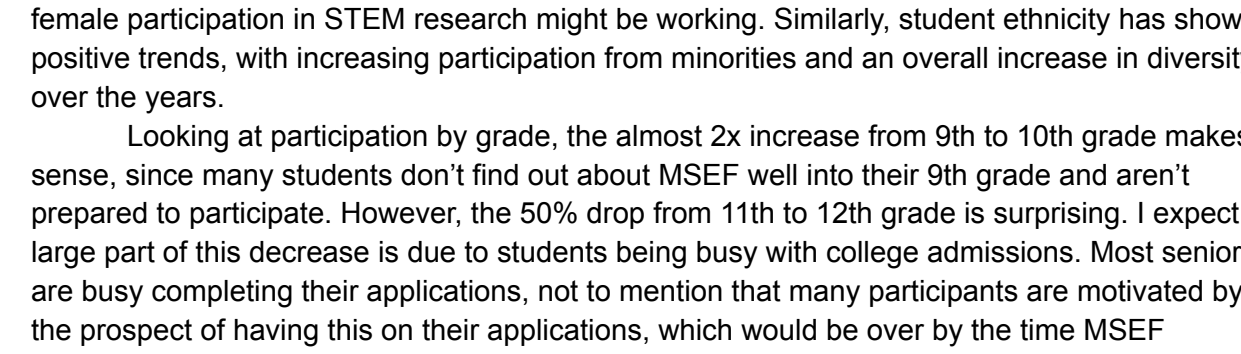


Fig 4.1: Plot of various AI/ML related search terms in Google search over the same time period shows the gain in popularity was not unique to MSEF data . Source: Google Trends.

For projects relating to cancer, there was a spike in 2017, possibly due to the Moonshot program that year (Singer, D. S. et. al. 2016). The frequency of cancer projects fell in 2018, but has been steadily increasing ever since. Similarly, projects related to infectious diseases increased from 2020-21 following the emergence of the COVID-19 virus.

4.2 Biases in Research Topics:

While there are trends in overall frequency of project topics, numerous factors were found to affect participant choice of project topic, including gender, grade, and potential impact. Fields like Engineering and Computer Science were heavily male-dominated, with other fields such as Biology, Environmental Science, and Social Science being female-dominated. This gendered preference in project topics is corroborated by other studies (Makarova, E. et. al. 2019), indicating that this bias is not just unique to MSEF. Certain project topics were also more common among grades, and overall there seemed to be a trend towards high-impact or trendy topics, in hopes of winning awards.

4.3 Factors Affecting Performance

Multiple factors impacted project performance, such as project topic, applied value/impact, team size, participant gender and ethnicity, and grade. A consistent trend throughout all 8 years is that projects in fields such as Computer Science, Biology, Environment or Behavioral/Social have the highest probabilities of being awarded. More specifically, high-impact fields like Climate Change, Healthcare and Environmental Science tend to win awards. Conversely, projects in pure Chemistry, Math or Physics, with little applied value typically perform poorly. Furthermore, inclusion of trending fields or topics can increase your probability of winning an award, with projects involving AI or Machine Learning seeing a significant increase in award probability. Both the number of projects and this award boost has been increasing, as AI/ML becomes more relevant in scientific research. Similarly, Cancer-related projects got a 30% boost in probability of placing in the fair, with 2017 seeing a huge spike in the number of awards. This fits the overall trend of fundamental versus applied research, where projects in basic science have a much lower probability of winning an award while high-impact research performs better. One would think that at this early stage where most students are still exploring various topics in STEM, developing strong foundations in the fundamentals should be important and that MSEF would encourage and reward fundamental research. In fact, applied research projects can be prone to unrealistic claims and exaggerations about the implications of the project, whereas fundamental research relies mainly on the scientific effort as presented. Numerous studies have shown that fundamental STEM research has been very beneficial to the economy (Salter, A. 2000).

Surprisingly, project title has an apparent impact on awards, as shown by the semantic clustering analysis, where some title clusters were enriched for awards and others had very few awards. There is a confounding factor here, as projects in the same fields are likely to have similar titles, however this still shows that even an inclusion of key terms in the project title can have an effect on how it performs in the fair. In the future, I want to try other more advanced clustering algorithms, as this current approach was very dependent on empirically determined weights, and could be tuned to yield more accurate groupings.

4.4 Participation and Performance of Schools:

The vast majority of schools submit very few projects, with a small set of outlier schools submitting hundreds of projects. This implies that in a few schools, MSEF participation might be heavily encouraged while in most of the other schools there is only occasional participation. This shows that schools themselves can have a large influence on interest in STEM and participation in MSEF. Also, as expected, there is a significant correlation between the participation of schools and the size of the school (total number of students). This highlights an opportunity for some of the larger schools to increase their representation in MSEF. On the other hand, there is no correlation between participation and per-student funding. This suggests that STEM interest can be achieved more easily through means that don't require significant monetary investment on part of the schools (Lichtenberger, E. and George-Jackson, C. 2013).

Surprisingly, the only factors that had a marginal effect on performance were academic ranking and school size. Higher ranked schools tended to win more awards, implying that high-ranked schools that don't participate in MSEF have the opportunity to excel in the fair and promote STEM research. There was a large cluster of these schools that have high academic ranking but low MSEF performance (See supplement 6.2).

4.5 Conclusions and Recommendations:

Overall, MSEF is having a positive impact on the Massachusetts student community. Participant diversity and participation from girls is steadily increasing, and there is little evidence of any negative biases or trends in participation although there is a slight overall drop in participation. The analysis highlights some recommendations for MSEF:

- Act on increasing participation and reverse the declining trend. Get more schools to participate in MSEF. Reduce the huge disparity seen in participation from schools where a few schools now account for a large number of the projects submitted.
- Encourage more fundamental research. Perhaps shift selection criteria to consider other factors such as effort, and scientific insight. Also, MSEF judging criteria could potentially be tweaked to reduce the importance of project title in performance in the competition.
- There is opportunity for some high academically ranking schools to do better in MSEF, those schools below the regression line could potentially do better. Similarly, some of the larger schools that fall below the regression line in participation could rethink their strategy and adopt programs to encourage more participation from their students.
- Reduce gender bias in choice of research topics.

Lastly, it is recommended that MSEF continue to repeat this analysis in future years, including more comprehensive data on the projects (lab use, continuation), student demographics, and data from regionals etc. I also intend to conduct additional analysis that might either corroborate or contest my conclusions. Either way, continuously monitoring these trends is the first step towards identifying potential improvements and striving towards a larger and more diverse STEM community.

5.0 Acknowledgements:

I would like to thank Shannon Gmyrek and Rebekah Stendahl for guidance and support on this project and the MSEF organization for providing the dataset and the internship opportunity.
I also would like to thank the other summer interns that worked on the project for initial discussions on this topic. I also would like to thank my family for their support and for reviewing and commenting on this report.

6.0 Supplementary Information:

Additional results from the clustering analysis and all Python and R scripts (Jupyter notebooks) used for the analysis are available at: https://github.com/rohan2017/MSEF_Internship
Data used for this analysis can be made available upon request and approval from MSEF.

6.2 List of high-ranked schools below the line (Fig. 3.4.3)

1. 'Newburyport High School'
2. 'Norwell High School'
3. 'Pioneer Charter School of Science'
4. 'Sturgis Charter Public School'
5. 'Braintree High School'
6. 'Advanced Math and Science Academy Charter School'
7. 'Algonquin Regional High School'
8. 'Hopedale Junior Senior High School'
9. 'King Philip Regional High School'
10. 'Marblehead High School'
11. 'Wellesley High School'
12. 'Bromfield School'
13. 'Cohasset Jr. Sr. High School'
14. 'Hamilton-Wenham Regional High School'
15. 'Andover High School'
16. 'Westford Academy'
17. 'Lincoln-Sudbury Reg. High School'
18. 'Winchester High School'
19. 'Canton High School'
20. 'Community Charter School of Cambridge'
21. 'Hopkins Academy'
22. 'Natick High School'
23. 'Tantasqua Reg. High School'
24. 'Brookline High School'
25. 'Concord-Carlisle Regional High School'
26. 'Wayland High School'
27. 'Blackstone Valley Voc-Tech. High School'
28. 'Chelmsford High School'
29. 'Medfield Senior High School'
30. 'O\'Bryant School Math/Science'
31. 'Manchester Essex Regional High School'
32. 'Arlington High'

7.0 References:

Agus, D. B. et. al., 2021. Cancer Moon shot 2.0. www.thelancet.com/oncology. Vol 22. https://doi.org/10.1016/S1470-2045(21)00003-6

Allahyari, M. et. al. 2017. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. KDD Bigdas, August 2017, Halifax, Canada

Becker, R., 2021. Water shortages: Why some Californians are running out in 2021 and others aren't. Calmatters.org.

Chablani, M. 2019. Semantic similarity Classifier and clustering sentences based on semantic similarity. Medium: Towards data science. https://towardsdatascience.com/semantic-similarity-classifier-and-clustering-sentences-based-on-semantic-similarity-a5a564e22304

Corbett, C. and Hill, C., 2015. Solving the Equation: Variables for Women's success in Engineering and Computing. www.aauw.org. ISBN: 978-1-879922-45-7

Diperi, D., 2021. The great resignation of women in STEM. WEST Wisdom Blog. http://info.westorg.org/blog/the-great-resignation-of-women-in-stem

Fry et. al. 2021., Stem jobs see uneven progress in increasing gender racial and ethnic diversity. Pew research center newsletter. https://www.pewresearch.org/science/2021/04/01/stem-jobs-see-uneven-progress-in-increasing-gender-racial-and-ethnic-diversity/

Gonsalez, H., 2012. An Analysis of STEM education funding at the NSF: Trends and Policy discussion. CRS report for Congress. https://www.commoncorediva.com/wp-content/uploads/2015/05/stemanalysis.pdf

Hornsey et. al. 2018. The Psychological Roots of Anti-Vaccination Attitudes: A 24-Nation Investigation., Health Psychology, Vol. 37, No. 4, 307–315 https://www.apa.org/pubs/journals/releases/hea-hea0000586.pdf

Karampelas, K. 2021., Trends on Science Education Research Topics in Education Journals. European Journal of Science and Mathematics Education https://www.scimath.net Vol. 9, No. 1, 2021, 1-12

Koch, K., 2020. A friendly introduction to text clustering. Medium: Towards Data Science. https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04.

Kong et. al., 2020. Reducing gender bias in STEM., MIT Science Policy Review. Vol1. Pg. 55

Kulmanov, M. et. al. 2021. Semantic similarities and Machine Learning with Ontologies. Briefings in Bioinformatics, 22(4), 2021, 1–18. https://doi.org/10.1093/bib/bbaa199

Lichtenberger, E. and George-Jackson, C. Predicting High School Students' Interest in Majoring in a STEM Field: Insight into High School Students' Postsecondary Plans - Journal of Career and Technical Education, Vol. 28, No.1, Winter, 2013

Makarova, E. et. al. 2019. The Gender Gap in STEM Fields: The Impact of the Gender Stereotype of Math and Science on Secondary Students' Career Aspirations. Front. Educ., 10 July 2019 | https://doi.org/10.3389/feduc.2019.00060 https://www.frontiersin.org/articles/10.3389/feduc.2019.00060/full

Potvin, P. and Hasney, A., 2014. Analysis of the Decline in Interest Towards School Science and Technology from Grades 5 Through 11. J Sci Educ Technol (2014) 23:784–802.

Riegle-Crumb, C., et. al. 2019. Does STEM Stand Out? Examining Racial/Ethnic Gaps in Persistence Across Postsecondary Fields. Educational Researcher, Vol. 48 No. 3, pp. 133–144 DOI: 10.3102/0013189X19831006

Robnett, R.D., 2015. Gender Bias in STEM Fields: Variation in Prevalence and Links to STEM Self-Concept. Psychology of Women Quarterly. https://doi.org/10.1177/0361684315596162.

Singer, D. S. Et.et. al. 2016. A U.S. Cancer Moonshot" to accelerate cancer research Science, Vol 353, Issue 6304

Souvaine, D.L. et. al. 2020. State of US Science and Engineering: Science and Engineering Indicators. Report by the National Science Board. NSF.

Zhan and Serwer., 2021. Chinese tech investment poses real danger to US industry., Yahoo Finance article. https://news.yahoo.com/chinese-tech-investment-poses-real-danger-to-us-industry-michael-dell-135530260.html#:~:text=Despite%20longstanding%20U.S.%20concern%20over,escalating%20competition%20with%20U.S.%20firms.

2020 Doctorate recipients from US universities. Nov. 2021 report. National Center for Science and Engineering Statistics. Directorate for Social, Behavioral and Economic Sciences. National Science Foundation. NSF 22-300

2020 Science and Technology: Public Attitudes, Knowledge and Interest. NSB-2020-7 National Science Board Indicators.