Data-Driven Investigation of the Role of Educational Resources in Shaping Science Fair Projects

Daniel Verdi do Amarante
daniel.verdidoamarante@richmond.edu
University of Richmond
Richmond, VA, USA

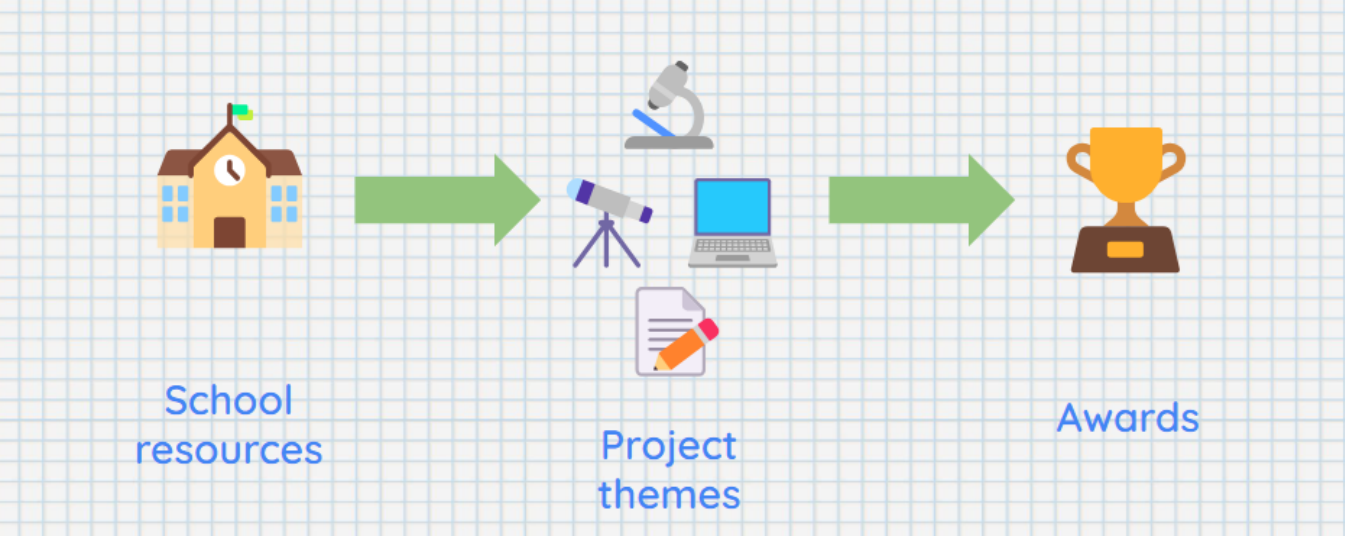

Figure 1: Simplified visual representation of the study; arrows indicate the relationships being examined (created by the author with background image from Freepik)

ABSTRACT

Science fairs provide opportunities for K-12 students to engage in scientific research from an early age. This study investigates the influence of school characteristics on the diversity and complexity of project topics, as well as the likelihood of different themes receiving awards in Brazilian science fairs. We scraped public data from the websites of the two major science and engineering fairs (FEBRACE and Mostratec) over a five-year period (2015-2019), and combined that with school data from a governmental agency (INEP). Using a combination of topic modeling and statistical analysis, we aim to explore the impact of educational resources on project themes and award outcomes. We hypothesize that schools with more resources will exhibit a greater diversity and complexity of project topics and higher award rates. Furthermore, we expect to find significant correlations between project themes and awards, which might be influenced by external factors such as judges interests and societal trends. Finally, we also investigate the connections between project topics and other variables, such as student gender and project category. This research contributes to a deeper understanding of the factors shaping student participation and success in science fairs, highlighting the importance of equitable access to resources and opportunities in fostering youth-driven scientific research. Results can bring important insights to understand and improve the science education process and formulation of public policies in the area.

KDD-UC '24, Barcelona, Spain,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Computing methodologies** → *Topic modeling*; • **Applied computing** → **Education**; • **Social and professional topics** → *Informal education.*

KEYWORDS

Topic modeling, STEM education, Equity

ACM Reference Format:
Daniel Verdi do Amarante. 2024. Data-Driven Investigation of the Role of Educational Resources in Shaping Science Fair Projects. In *Proceedings of Undergraduate Consortium at the International Conference on Knowledge Discovery and Data Mining, August 25–20, 2024 (KDD-UC '24).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Science fairs are critical incubators for young researchers, encouraging innovation and creativity through hands-on experience and mentorship. In Brazil, these events play a pivotal role in promoting scientific literacy and encouraging the next generation of scientists. Additionally, many science fairs are instruments of public policies in Brazil, receiving federal funding and benefitting thousands of students yearly.

However, despite their importance, there has been limited systematic, quantitative analysis of the thematic trends and intellectual diversity present in these student projects. There's only one research to date that has investigated the themes of projects in science fairs, which explored how topics have changed over time and across regions [5]. By using topic modeling techniques, the study revealed the variations in project topics over time and across different regions and school settings in Brazil, showcasing how topic modeling can uncover trends and patterns in student research.

Also making use of topic modeling, but for another educational context, another study examined the relationship between household income and college admission essay content in the United States [2]. They discovered that essay content and style were more strongly correlated with household income than SAT scores, highlighting how socio-economic factors significantly influence non-numerical educational assessments.

On another line of research, some studies have explored the impact of socio-economic and demographic factors on the performance of high school students in science fairs. For instance, one study found that school type and English proficiency significantly influenced success in a Brazilian science fair, showing the need to address systemic inequalities to ensure fair competition and equitable opportunities [11].

These studies collectively illustrate the potential of data science techniques to reveal important insights into educational outcomes and disparities. At the same time, they reveal a lack of understanding the connections between socio-demographic factors and the project topics in science fairs. This knowledge gap impedes our ability to better assess the evolving interests of young Brazilian scientists and to guide policy and educational strategies effectively.

Addressing that problem and combining the methodologies of these studies, our research aims to understand how resource availability might shape the thematic choices of student projects in science fairs, and how those themes relate to the events' awards. Our analysis focuses on two primary research questions:

- **What is the impact of school characteristics on project topics?** We hypothesize that the availability of resources at a school significantly influences the diversity and complexity of the research topics chosen by students. Specifically, schools with greater access to scientific and pedagogical resources, such as science and computer labs, will have students who select a broader range of more innovative and complex project topics.
- **What is the relationship between project topics and awards received?** We hypothesize that certain project themes or types of projects are more likely to receive awards than others, potentially due to their alignment with judges' interests, societal trends, or the perceived importance and impact of the research.

Our analysis bridges the gap on the current state of student-led scientific inquiry in Brazil, but can also offer valuable insights for educators, policymakers, and stakeholders looking to nurture and support young talent. Through this work, we hope to contribute to the broader discourse on science education and the development of young researchers, both in Brazil and globally.

2 RELATED WORK

Although there are not many studies examining science fairs with large-scale, quantitative methods, research on educational outcomes and the factors influencing them is extensive. Various studies have examined how socioeconomic and institutional characteristics impact student achievements across different contexts, mainly in college admissions.

For instance, it has been demonstrated that socioeconomic status (SES) significantly influenced standardized test scores and college attendance rates, and that controlling for SES dramatically altered school rankings [10]. Another study also found that family income significantly influences SAT performance, with the effects being substantial and non-linear, and nearly twice as large for Black students compared to White students, even when accounting for parental education and high school achievement using structural equation modeling [4].

Investigating another educational context, other researchers revealed that socio-economic advantage and urban locations positively influenced computer science uptake in Scottish secondary schools [8]. All these studies highlight the importance of accounting for socio-economic factors when evaluating educational performance, a principle that underpins our research.

There is also substantive research on the production of science. Since science fairs are a culmination of a process that is both educational and scientific, it is crucial to also consider that literature. The "diversity paradox" in scientific innovation, for instance, has been recently explore [7]. The researchers found that underrepresented groups produce higher rates of scientific novelty, but their contributions are often devalued and discounted. This study highlights systemic biases that can affect recognition and success, informing our consideration of how school resources and demographic factors might influence the thematic choices and success rates of student projects.

Various studies have applied topic modeling techniques to study research trends, demonstrating the utility of topic modeling in predicting and uncovering trends in large research datasets [1, 3, 9]. Their methodologies provide a framework for our approach to analyzing science fair projects, allowing us to understand scientific themes explored by the student scientists.

3 METHODOLOGY

3.1 Data

The data for this research were collected from the two major national science fairs in Brazil: Mostra Internacional de Ciência e Tecnologia (Mostratec), established in 1985, and Feira Brasileira

de Ciências e Engenharia (FEBRACE), established in 2003. We focused on a recent five-year period from 2015 to 2019, intentionally avoiding the pandemic years that likely had different project dynamics. The data was scraped from the fairs' official websites using Python scripts, where project information was made available in PDF format.

We collected detailed information from the events' proceedings available on the FEBRACE and Mostratec websites. The scraped data includes the project abstracts, authors names (students and mentors), schools names, cities, and states, scientific field, and awards. The final dataset encompasses information from 3,317 projects (1,694 from FEBRACE and 1,623 from Mostratec), representing almost 900 schools across all states in Brazil.

To complement the project data, we obtained additional information about the schools from the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), the National Institute for Educational Studies and Research. This dataset includes:

- Facilities: Binary indicators for the presence of a science lab, computer lab, library, internet access, and scientific-pedagogic materials.
- Entrance Exam: Binary indicator of whether the school requires an entrance exam for admission.
- School Location Type: Classification of the school as either rural or urban.
- School Administrative Type: Classification of the school as private, city-administered, state-administered, or federally administered.
- Teacher-Student Ratio: The number of teachers per student at each school.

3.1.1 Data Processing. We inferred the gender of the project authors at the project level (categorized as Male, Female, or Mixed) using an R package specifically designed for Brazilian names, genderBR. We recognize the limitations of this approach, such as inaccuracies in gender prediction for ambiguous or uncommon names and the inability to account for non-binary or gender-fluid identities. However, we used this method because it provides a scalable and systematic way to incorporate gender information into our analysis, which can help understand potential gender disparities in science fair participation and outcomes.

Since Mostratec also receives projects from other countries, we excluded rows for projects that were not from Brazil, as we would not have the school data for those from INEP.

The datasets obtained from the science fairs' websites were linked to the INEP datasets by using a fuzzy matching algorithm, and then manually checked by the author. Since the school names were typed by the participants when registering to the fairs, names were not standardized. It was then necessary to make all names lower case, use the stringdist library in R to do the fuzzy matching (considering only school names in the project city), and then check each of them manually. While the automatic matching process worked for most of the schools, a few of them were misclassified, so verifying it manually was essential for the quality of the analysis.

3.1.2 Ethical Considerations. This research received an Institutional Review Board (IRB) exemption as it involves the collection and analysis of existing, publicly available data. The data presented on the science fairs' websites was provided by the participants (students and teachers), who signed consent forms agreeing to the public release of their project abstracts and related information.

3.2 Topic Modeling

Following standard practices, we pre-processed the textual data before applying the topic models. This was done by removing stop words, punctuation, and numbers, lower-casing all characters, lemmatizing words. This is the current state of the research, finalizing data cleaning and preparation.

To analyze the themes of science fair projects, we plan to employ BERTopic, a recent deep learning model for topic modeling [6]. The number of topics will be selected using coherence scores to ensure optimal topic separation and interpretability. Once the model is trained, we will analyze the topics by examining the most representative terms and manually labeling the topics to reflect the themes.

3.3 Statistical Analysis

To explore the relationships between school characteristics, project topics, and awards, we will conduct a series of statistical analyses. First, we will integrate the topic modeling results with the school characteristics and award data to form a comprehensive dataset. Then, we will perform some exploratory data analysis and calculate descriptive statistics to summarize the distribution of school characteristics, project topics, and awards. Then, we will employ a linear regression model to assess the impact of school characteristics on project topics, and a logistic regression to model the likelihood of winning awards based on school characteristics and inferred project topics.

4 PRELIMINARY AND EXPECTED RESULTS

While the analysis is still in progress, we describe here some of the anticipated key findings. We expect to find that schools with more resources, such as science labs, computer labs, libraries, internet access, and a higher teacher-to-student ratio, will be associated with more complex and resource-intensive project topics. For example, students from well-resourced schools might undertake projects involving advanced experimental procedures, such as genetic engineering or robotics, which require access to specialized equipment and materials. Conversely, students from under-resourced schools might focus on less resource-intensive projects, such as environmental studies using local data or theoretical computer science projects. However, initial analysis of the data indicated little variation in the binary variables about school resources among the projects, as shown in Table 1. This might be because of the level of the science fairs analyzed here, which are nationally recognized and require an admission process. Therefore, it might be the case that only the projects that had access to resources made it to participate in these national events. Further analysis into the distribution of the variables in the overall schools in Brazil will be done, along with a search for other possible variables that could be used as proxies for access to resources.

Besides school resources, we also expect to find connections between other demographic variables and the topics explored by

Table 1: Distribution of projects for school binary variables

Variable	Existent in the school	Inexistent	No data
Entrance exam	1,705	1,383	229
Scientific materials	2,058	1,030	229
Science lab	2,540	548	229
Computing lab	2,818	270	229
Library	2,838	250	229
Internet access	2,970	118	229

the students. For example, students from rural schools might engage more in projects that use locally available materials and data, such as studies on local water quality or agricultural techniques. With gender inferred from project authors' names, we might also discover gender-related associations in the choice of project topics. For instance, female students might be more present in projects in the biological sciences or social sciences due to societal stereotypes or barriers to access other fields. Initial analysis of gender by student-indicated field, as shown in Figure **??**, revealed some of those patterns, so further exploration can bring interesting insights. Identifying these patterns can help demonstrate how gender dynamics influence STEM education and recognition, and possibly where interventions can be made.

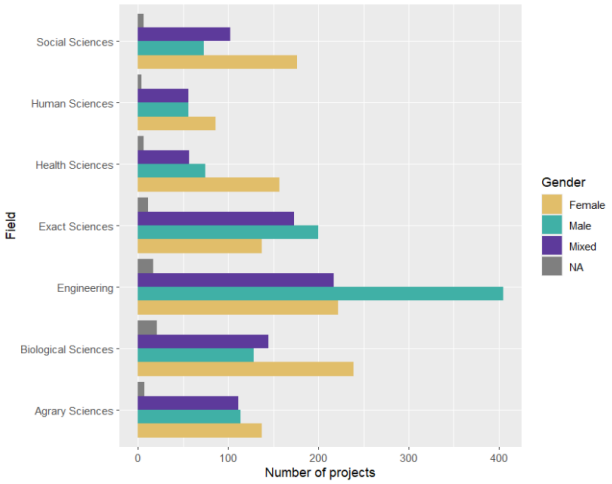

Figure 2: Inferred students' gender per project by field

Furthermore, we hypothesize that certain project topics will be more likely to receive awards than others, potentially due to their alignment with contemporary scientific trends, societal importance, or the interests of the judges. For instance, projects related to renewable energy solutions, such as solar-powered devices or wind energy innovations, may receive higher recognition due to the global emphasis on sustainability. In general, we also expect that applied science and engineering projects get more recognized due to their direct societal impact when compared to basic research. By analyzing the award distribution across different themes identified by BERTopic, we can identify which types of projects are most favored in these science fairs.

Finally, we will also explore how the fields indicated by students align with the themes identified through topic modeling. For instance, a student-indicated field of "Environmental Science" might align with BERTopic themes such as "Climate Change Mitigation" or "Sustainable Agriculture." Additionally, we might be able to observe preferences towards choosing a field based on the projects' application and methods. A project applying a machine learning algorithm to improve cancer diagnosis, for example, could be categorized as either "Health Sciences" or "Computer Science." By comparing the self-reported fields with the themes derived from text analysis, we can assess the accuracy and granularity of students' categorizations and understand how students' perceptions of their work match broader scientific trends. This analysis will help us identify any gaps or overlaps in how students classify their projects versus the underlying thematic structure revealed by topic modeling.

5 DISCUSSION AND CONCLUSION

Our research makes important contributions to the field of science education by integrating data science techniques to understand the relationships between school characteristics, project topics, and awards in science fairs. However, some limitations must be acknowledged. Two of the major ones are the reliance on publicly available data, which may have restricted the scope of our analysis, and the sample size of our study, which can limit the generalizability of our findings.

Moving forward, future research can expand the scope of data collection to include a broader range of science fairs and years, as well as complement the results with qualitative interviews or surveys with students, teachers, and judges. Moreover, longitudinal studies tracking students' participation and success in science fairs over time would offer a more comprehensive understanding of the long-term impacts of school resources and project topics on student outcomes.

Our study aims to provide a deeper understanding of how educational resources shape the intellectual landscape of student research in K-12 education. We hope to contribute valuable insights into the systemic factors that drive innovation and success in science fairs, informing educational policy and resource allocation strategies. By revealing the critical the role of school resources and possible disparities, we aspire to promoting greater equity and excellence in STEM education, encouraging diverse and high-quality scientific research among students.

ACKNOWLEDGMENTS

Thanks to my research mentor on this project, Dr. Lilla Orr, to Caleb Kwakye for helping with the data collection process, and the University of Richmond for the ongoing support.

A generative AI tool (ChatGPT) was used to create and edit parts of this manuscript.

REFERENCES

[1] Tesfamariam M Abuhay, Yemisrach G Nigatie, and Sergey V Kovalchuk. 2018. Towards predicting trend of scientific research topics using topic modeling. *Procedia Computer Science* 136 (2018), 304–310.

[2] AJ Alvero, Sonia Giebel, Ben Gebre-Medhin, Anthony Lising Antonio, Mitchell L Stevens, and Benjamin W Domingue. 2021. Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications. *Science advances* 7, 42 (2021), eabi9031.

[3] Stijn Daenekindt and Jeroen Huisman. 2020. Mapping the scattered field of research on higher education. A correlated topic model of 17,000 articles, 1991–2018. *Higher Education* 80, 3 (2020), 571–587.
[4] Ezekiel J Dixon-Román, Howard T Everson, and John J McArdle. 2013. Race, poverty and SAT scores: Modeling the influences of family income on black and white high school students' SAT performance. *Teachers College Record* 115, 4 (2013), 1–33.
[5] Adelmo Eloy, Thomas Palmeira Ferraz, Fellip Silva Alves, and Roseli de Deus Lopes. 2023. Science and engineering for what? A large-scale analysis of students' projects in science fairs. *arXiv preprint arXiv:2308.02962* (2023).
[6] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).
[7] Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. 2020. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences* 117, 17 (2020), 9284–9291.
[8] Fiona McNeill, Blaga Baycheva, Aba-Sah Dadzie, and Eleanor Mitchell. 2023. Exploring the Impact of School Location on Young People's Likelihood of Studying Computing in Scotland. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*. 389–395.
[9] Michael Paul and Roxana Girju. 2009. Topic modeling of research fields: An interdisciplinary perspective. In *Proceedings of the International Conference RANLP-2009*. 337–342.
[10] Robert K Toutkoushian and Taylor Curtis. 2005. Effects of socioeconomic factors on public high school outcomes and rankings. *The Journal of Educational Research* 98, 5 (2005), 259–271.
[11] Daniel Verdi do Amarante and Kritim Rijal. 2024. Statistical Analysis of Socio-Economic Factors as Predictors of Performance in a Brazilian Science Fair. *Revista Educação Pública: Divulgação Científica e Ensino de Ciências* (2024).

Received 27 May 2024; accepted 19 June 2024