What Factors Correlate with the Relationship Between Gender and Race and Pursuing STEM?

Saket Reddy

10/16/22

Abstract

The world of technology and Science, Technology, Engineering, and Math (STEM) is dominated by Caucasians, Asians, and males. According to Computer Science.org, while African Americans and Latinos make up 13.4% and 16% of the US population respectively, they only comprise 9% and 7% of STEM occupations (McGee). In this paper, I sought to answer the question, "How do races and genders differ in the way they pursue STEM, and what factors correlate with these differences?" Identifying which factors cause this lack of representation is the most important step in fixing the diversity problem.

To answer this question, I used data from the High School Longitudinal Study (HSLS), which asked hundreds of questions at different points throughout students' high school and college careers. I used 16 of these questions and analyzed how strongly they correlate with pursuing a STEM career or not pursuing a STEM career. I found that Asians are much more likely to pursue STEM, African Americans are much less likely to pursue STEM, and women are much more likely to pursue STEM, although they largely pursued health occupations within STEM.

It is important that we make sure everyone feels comfortable pursuing or not pursuing STEM. From my literature review, I found that Asians are more likely to and African Americans are less likely to pursue STEM because Asians feel like they fit in and African Americans feel like they don't (Milgrom-Elcott). In order to increase the diversity in the STEM workforce, we need to make sure all minority groups feel welcomed in high school STEM classes, college STEM courses, and STEM workplaces themselves.

1. Introduction

The STEM workforce is noticeably lacking in diversity. As mentioned earlier, according to Computer Science.org, while African Americans and Latinos make up 13.4% and 16% of the US population respectively, they only comprise 9% and 7% of STEM occupations (McGee). Additionally, women only make up 25% of computer jobs and 14% of engineering jobs (McGee).

The key to solving this problem is to answer the question, "How do races and genders differ in the way they pursue STEM, and what factors correlate with these differences?" By answering this question, we will be able to start working on removing the barriers or obstacles that minority students may face when pursuing STEM, or we might be able to make the STEM field look more welcoming to minority students. If, for example, we saw most people in a certain minority group strongly share a factor associated with not pursuing STEM, such as not taking math classes because it would not leave enough time for extracurriculars, we would see that we should work on addressing that factor.

I used a logistic regression model in order to classify the students as pursuing STEM or not pursuing STEM according to what they answered on each of the 16 chosen survey questions. I made a model for all of the students, and different models for each of the following ethnicities: White, Asian, African American, and Latino. All of the data was categorical. The main output I analyzed were the correlation coefficients, which showed how strongly each factor/question correlated with pursuing STEM (a coefficient from 0 to 1) or not pursuing STEM (a coefficient from 0 to -1).

2. Background

There have been several studies conducted and papers published related to pursuing STEM, and how to make the STEM workplace more diverse. One example is "A Gender Analysis of the Occupational Pathways of STEM Graduates in Canada" (Frank). This study analyzed survey responses of men and women with STEM credentials, using a longitudinal survey conducted in 2006 and 2016, to see what occupations they end up in. They found that, in general, men with STEM degrees were usually more likely to work in a STEM occupation, while women with a STEM degree and job were more likely to switch to a non-STEM occupation. They also found other interesting results, such as that men who studied computer or information science were less likely to leave a STEM occupation than men who studied engineering, and younger STEM graduates were more likely to leave a STEM job than older ones.

One advantage which this study has was that it analyzed many questions, and analyzed them with more depth. Not only did the study look into the difference between adult men and women in STEM occupations, but they also looked into socioeconomic, educational, age, and wage differences. This study also had the advantage of analyzing people who already had STEM degrees and careers, whereas the HSLS data I am using asked questions to people who did not have either of these. That being said, because this study analyzed adults between the ages of 25 and 54, it misses out on looking into career aspirations at a younger, potentially more important age.

Another research paper I found is "Factors Influencing STEM Career Aspirations of Underrepresented High School Students" (J. Mau and Li). This study used HSLS data to look at many different factors that may be associated with minority students' career aspirations. They found that family support, school influence, and self-esteem in STEM were important for choosing a STEM career. They also found that self-esteem in STEM was the most important factor for choosing a STEM career, and students who chose STEM careers usually felt less school belonging. One advantage of this study is that they looked at more factors than I did. Additionally, they split their factors into different "clusters," which were based on a model by Mau and Bikos (2000) made with extensive research.

One disadvantage of this study is that the dependent variable they used was a question which asked ninth graders what job they wanted to have at age 30. As they mentioned in their paper, this may have led to misleading conclusions since ninth graders may not be mature or

knowledgeable enough to know what occupation they want to have at age 30. In contrast, my study's dependent variable was a question asking students what job they wanted to have at 30 in 2016, a time when most students were in college.

3. Dataset

I used data from the High School Longitudinal Study (HSLS), which asked hundreds of questions at different points throughout students' high school and college careers. The survey started in 2009 when the students were ninth graders, and the most recent survey was conducted in 2016, when many students were in college. I used 16 of the questions asked during the high school surveys and analyzed how strongly they correlate with pursuing a STEM career or not pursuing a STEM career. All of the data was categorical. I made a model for all of the students, and different models for each of the following ethnicities: White, Asian, African American, and Latino. I utilized logistic regression models where 80% of each model's samples were saved for training data.

The outcome variable (called "-age_30_job" in this study) is a question asking students what occupation they wanted to have at age 30. This question was asked multiple times throughout the students' education careers, but I chose the question that was asked in 2016, when most students were in college, so that my model could be predictive over time. All other survey questions were asked when students were in high school.

Table 1. Shows the number of students within each model.

Model	**Number of Samples**
General (all races)	9232
White	5194
Asian	820
African American	977
Latino	1375

Below, variable names that start with "-" are nominal (their options do not have an order) while all other variables are ordinal (their options have an order to them, such as "strongly agree", "agree", "disagree", "strongly disagree"). Most of the nominal variables used are binary (they only have two options).

For more information on each survey question, such as the answer choices for each question, please look at the codebook.

Codebook:
https://docs.google.com/document/d/130axeKBPW7I8q_PZDKsFOqn_DSOZFZqLG930tm_HiwU/edit?usp=sharing

Table 2. Shows which variable name corresponds with which survey question.

Survey Question	**Variable Name**
Race	-race
Gender	-gender
As far as you know, are the following statements true or false for your closest friend? Your closest friend gets good grades. [Asked in 9th grade]	-friend_good_grades
Do you plan to take more math classes because students like you do? [Asked in 9th grade]	-continue_bc_similar_people_do_math
If there were no barriers, how far in school would you want to go? [Asked in 11th grade]	how_far_wants_go
How much do you agree or disagree with the following statements about your teacher for [math course title]? Remember, none of your teachers or your principal will see any of the answers you provide. Your teacher [treats/treated] some kids better than other kids. [Asked in 11th grade]	teacher_does_not_treat_others_better
[Are you currently taking any science, computer science or technology courses/Were you taking any science, computer science or technology courses during the spring term of 2012? [Asked in 11th grade]	-taking_comp_sci
How much do you agree or disagree with the following statements? You see yourself as a math person. [Asked in 11th grade]	not_math_person
How much do you agree or disagree with the following statements? You see yourself as a science person. [Asked in 11th grade]	not_science_person
How much do you agree or disagree with the following statements? If you spend a lot of	stem_does_not_mean_no_activity_time

time and effort in your math and science classes, you won't have enough time for extracurricular activities. [Asked in 9th grade]	
How much do you agree or disagree with the following statements about your current school? Getting good grades in school is important to you. [Asked in 9th grade]	good_grades_not_important
How much do you agree or disagree with the following statements about the usefulness of your [fall 2009 math] course? What students learn in this course will be useful for a future career. [Asked in 9th grade]	math_course_not_useful_job
How much do you agree or disagree with the following statements about the usefulness of your [fall 2009 science] course? What students learn in this course will be useful for a future career. [Asked in 9th grade]	science_course_not_useful_job
How much do you agree or disagree with the following statements about your [fall 2009 science course]? You are enjoying this science class very much. [Asked in 9th grade]	not_enjoying_science_course
How important to [you/your teenager] [were/was] each of the following characteristics when choosing to attend [November 1 2013 postsecondary institution]? Academic quality or reputation. [Asked in 11th grade]	academic_quality_not_important_for_choosing_college
What job do you want to have at age 30? [Asked when most students were in college]	[Outcome variable] -age_30_job

Before using the data, I pre-processed the dataset. I began by narrowing down the hundreds of questions asked to only 16 questions. When narrowing down the questions, I prioritized using questions which were answered by many students and balancing the different types of questions, such as demographic, STEM-focused, and aspirational questions. Then, because all of these questions were categorical, I made sure all of the answers in the database were coded an integer to remove any samples with errors made when inputting the respondent's answers.

After this, I removed the students who answered any question in a way that couldn't be analyzed in the model. For example, if a student had an "answer" of -6, which means "component not

applicable," to the question asking them what they want to be at age 30, they would be removed since this can't be analyzed in a useful way.

This being said, for many of the questions used as independent variables, I did keep the students who "answered" -9, -8, and -7, which mean "missing", "unit non-response", and "item legitimate skip/NA," respectively. This is because I imputed these values with the most common response for that survey question. For example, for all of the students who "answered" -9, -8, or -7 for the question asking if their math teacher treated some students better than others, I replaced their answer with a 3 (which means "disagree"), because that was the most common answer to this question. I imputed the data because the original dataset was too small after I had removed all of the missing observations.

Next, I one-hot encoded the race factor. I made a variable for each ethnicity and made the students' "answers" a 1 for the ethnicity they answered on the survey and a 0 for all of the other ethnicities. Finally, I recoded the "-age_30_job" column so that it was a 1 if a student pursued any of the STEM subfields and a 0 if the student didn't pursue STEM or if their occupation was unreadable.

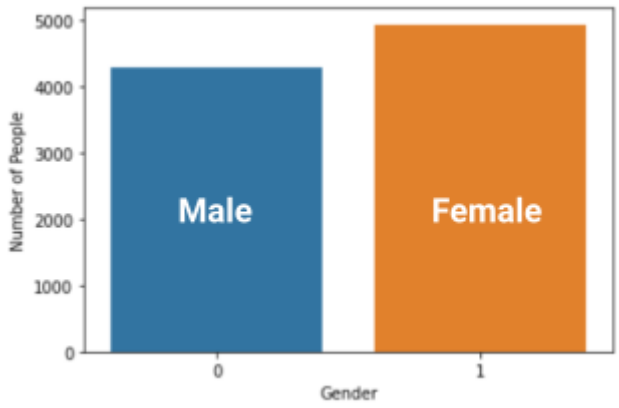

Fig 1. Graph depicting how many people of each sex were in the data.

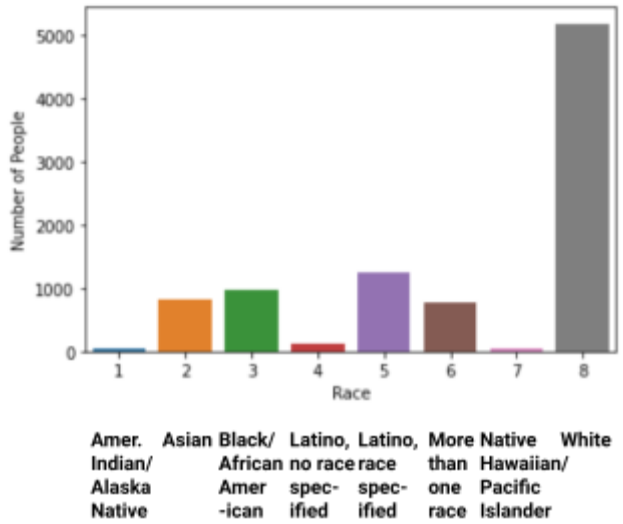

Fig 2. Graph depicting how many people of each race were in the data.

4. Methodology / Models

I wanted to use both a predictive and causal approach to this study. Therefore, I used logistic regression to predict whether or not a student will pursue STEM based on their responses to certain survey questions. Then, I analyzed the correlation coefficient for each factor (survey question) to decide which factors were most important for pursuing a STEM career, and did external research to find why these factors were important for pursuing a STEM career.

I used logistic regression, a classification algorithm, in order to analyze the data. Logistic regression works by looking at training data and analyzing how much each variable correlates with the outcome variable. Then, the model makes a classifier and uses this to predict the outcome of more data. This was an appropriate algorithm to use since I was trying to look at what features are most important for an output - whether or not a student will pursue STEM - and logistic regression automatically finds which features are most important so it can make the most accurate predictions. Using a linear regression model did not make sense because the outcome variable was categorical. I also did not use other classification models. For example, I wanted all of the variables to be considered holistically, so a decision tree was not appropriate to use.

All of the data I used was either nominal or ordinal, so I did not have to consider numerical data when making my models. When making my logistic regression models, I first removed the "stu_id" (Student ID) column, since it is not useful in classifying the data, and the "-race" column, since race is one-hot encoded. I also removed the columns for each race in the race-specific dataframes. After this, I prepared the logistic regression models for training by saving 80% of the data for training and sorting all of the variables into inputs and outputs (the only output was "-age_30_job"). Finally, I created the logistic regression models - one for all races and one for each race individually - and created a table of the correlation coefficients for

each feature sorted by strength. I created separate models for each race individually to see if there were any interesting observations that differed by race.

Each feature (survey question) has a correlation coefficient, which is a number in-between -1 and 1. The closer to 1 a feature is, the more strongly that feature is associated with pursuing STEM. In contrast, the closer to -1 a feature is, the more strongly that feature is associated with not pursuing STEM. Features whose coefficients are close to 0 have very little association with pursuing or not pursuing STEM.

5. Results and Discussion

My research findings came from analyzing the correlation coefficients between each survey question chosen and whether or not the person pursued a STEM career. Some of the correlations were expected. For example, I expected that "not_science_person" would be negative. Because these people do not identify themselves as "science people," they likely are not interested in pursuing a STEM career. However, some of the results were a surprise.

For example, being a woman was always strongly correlated with pursuing a STEM career. In fact, in the general model (hsls_data_logreg_model) with all races, women were twice as likely to pursue a STEM career. By far, the most popular category for women was health occupations, as shown by Fig. 4 below. Furthermore, most of the people who chose health occupations as their age 30 STEM career were women. Because health occupations were such a large category and this category was dominated by women, this is most likely why women were strongly correlated with pursuing a STEM career. I found that women may be more likely to pursue health occupations because women already have greater representation in the healthcare industry and women have greater career satisfaction on average in the healthcare industry (Berlin et al.).

Another surprising result that I would like to discuss is that my model showed me that women are more likely to want to go far in college, and academic reputation is important to them when it comes to choosing a college. This may be explained by the fact that women are more likely to pursue occupations such as nursing which usually require higher degrees, they may be more likely to want to attend a prestigious college and pursue a high degree. While many health-related occupations such as nursing are considered post-graduate degrees, men are more likely to pursue careers such as those in IT which don't require advanced degrees, so men may be more likely to want to go to college for less time.

One more interesting result is that being Asian is moderately correlated with pursuing a STEM career, with Asians being 1.516 times as likely as other races to pursue STEM. This may be because there are already so many Asians in STEM, so Asian want to pursue this field because they feel like they fit in. Additionally, an external article showed me that Asian immigrant parents tend to want their children to have good, well-paying jobs, such as careers in computer science and other STEM fields (Li et al.). Asian immigrants may also be more likely to pursue

STEM rather than the humanities since it is perceived that the humanities require a strong understanding of the English language.

In contrast, I found that being African American is moderately correlated with not pursuing a STEM career, with African Americans being 22.65% less likely to pursue STEM. This may be because there are already few African Americans in STEM careers, making it seem like African Americans don't fit in. African Americans see higher drop-out rates and college debts than other ethnicities, making it hard to pursue STEM (Lennon).

With all of this being said, there are several limitations to this study. Much of my data was heavily skewed. In addition to the outcome variable being mostly people pursuing health occupations or not pursuing STEM at all, the study was also dominated by Caucasians. My datasets were also relatively small, with my general model of all races having less than 10,000 people in it. This, combined with the fact that the study was dominated by Caucasians, meant that African American-only and Asian-only datasets had less than 1,000 people in them. These reasons are likely why my metrics, shown in Table 8 below, were relatively low.

Table 3. Shows the correlation coefficients for the general model (all races included):

not_science_person	-0.491171
african_american	-0.256837
not_math_person	-0.209538
native_american	-0.180277
science_course_not_useful_job	-0.156864
latino	-0.144861
multiracial	-0.109815
academic_quality_not_important_for_choosing_college	-0.098438
math_course_not_useful_job	-0.094436
white	-0.081019
good_grades_not_important	-0.069179
no-bio_adoptive_step-parent	0.000000
-continue_bc_similar_people_do_math	0.007713
stem_does_not_mean_no_activity_time	0.024837
native-hawaiian_pacific-islander	0.028566
teacher_does_not_treat_others_better	0.057893
-friend_good_grades	0.069125
not_enjoying_science_course	0.089728
how_far_wants_go	0.200865
-taking_comp_sci	0.259954
asian	0.416333
-gender	0.696382

Table 4. Shows the correlation coefficients for the model with only students who are white:

not_science_person	-0.506567
not_math_person	-0.222537
math_course_not_useful_job	-0.192557
academic_quality_not_important_for_choosing_college	-0.125443
science_course_not_useful_job	-0.111061
good_grades_not_important	-0.080706
-continue_bc_similar_people_do_math	-0.060589
-friend_good_grades	-0.026165
teacher_does_not_treat_others_better	0.025286
stem_does_not_mean_no_activity_time	0.066760
-taking_comp_sci	0.086144
not_enjoying_science_course	0.089486
how_far_wants_go	0.178337
-gender	0.772939

Table 5. Shows the correlation coefficients for the model with only students who are Asian:

not_science_person	-0.451731
-continue_bc_similar_people_do_math	-0.293896
not_math_person	-0.185312
math_course_not_useful_job	-0.124794
science_course_not_useful_job	-0.109603
good_grades_not_important	-0.108721
-friend_good_grades	-0.037944
teacher_does_not_treat_others_better	-0.031846
not_enjoying_science_course	0.017474
stem_does_not_mean_no_activity_time	0.094528
academic_quality_not_important_for_choosing_college	0.163427
-gender	0.254798
-taking_comp_sci	0.391358
how_far_wants_go	0.396570

Table 6. Shows the correlation coefficients for the model with only students who are African American:

not_science_person	-0.376102
-continue_bc_similar_people_do_math	-0.360345
science_course_not_useful_job	-0.357817
not_math_person	-0.100957
teacher_does_not_treat_others_better	-0.086676
not_enjoying_science_course	-0.084296
stem_does_not_mean_no_activity_time	0.044342
academic_quality_not_important_for_choosing_college	0.064829
good_grades_not_important	0.115921
-taking_comp_sci	0.181130
-friend_good_grades	0.196315
math_course_not_useful_job	0.257635
how_far_wants_go	0.337454
-gender	0.648654

Table 7. Shows the correlation coefficients for the model with only students who are Latino:

not_science_person	-0.423793
not_math_person	-0.369381
science_course_not_useful_job	-0.311748
academic_quality_not_important_for_choosing_college	-0.160529
-friend_good_grades	-0.151908
math_course_not_useful_job	-0.111723
stem_does_not_mean_no_activity_time	-0.099322
-taking_comp_sci	-0.027601
not_enjoying_science_course	0.033165
teacher_does_not_treat_others_better	0.037272
how_far_wants_go	0.139605
-continue_bc_similar_people_do_math	0.153572
good_grades_not_important	0.301615
-gender	0.871273

Table 8. Shows the metrics for each model:

	Accuracy	Precision	Recall	Hyperparameter C
General Model	0.67	0.63	0.42	1
Model with only students who are white	0.68	0.61	0.45	1
Model with only students who are Asian	0.65	0.67	0.80	1
Model with only students who are African American	0.69	0.65	0.30	1
Model with only students who are Latino	0.68	0.57	0.35	1

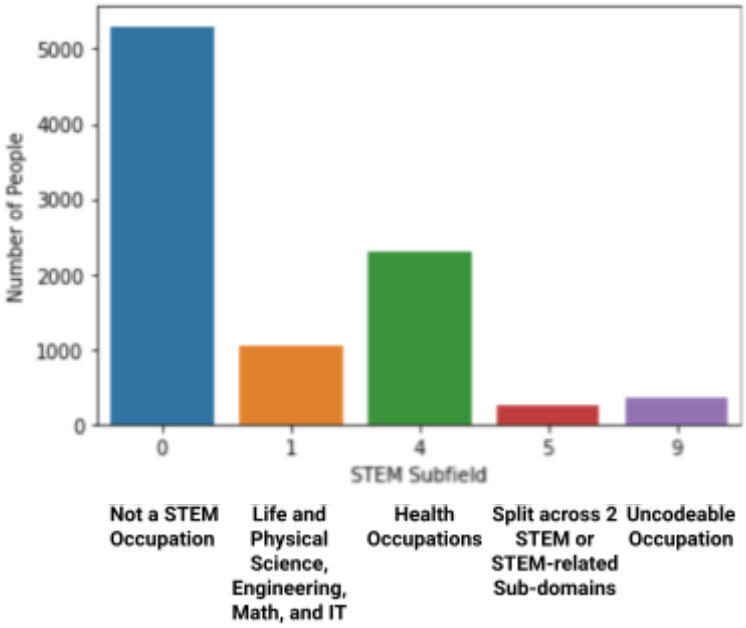

Fig 3. Graph depicting how many people chose each type of STEM job.

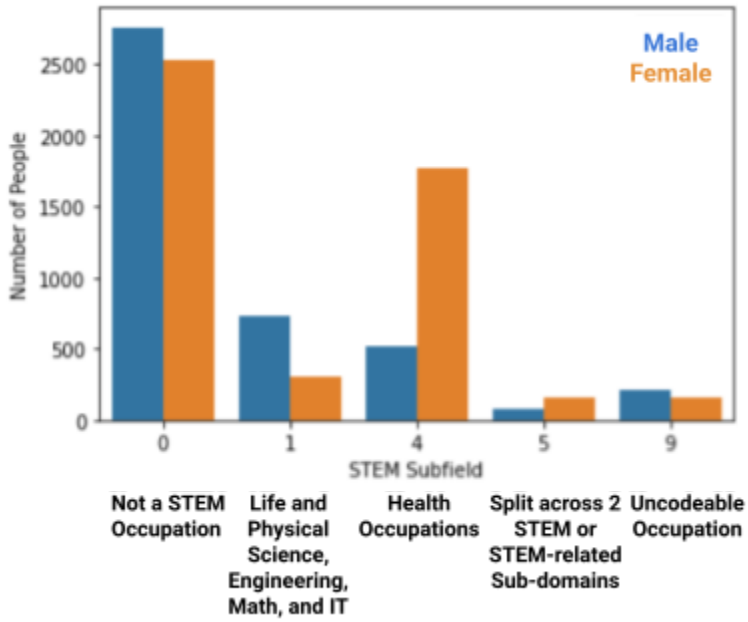

Fig 4. Graph displaying how many people of each gender chose each type of STEM job.

6. Conclusions

When I first started my research, my goal was to identify common trends in people who pursue STEM careers, especially trends by gender or race, in order to remove barriers for certains groups who are underrepresented in STEM. Through my research, I was able to identify such trends, although my model was less accurate than I would have preferred. I used logistic regression models to classify whether people would or would not pursue STEM depending on certain factors. However, my model performance was impacted by a small dataset size and using skewed data. Even so, I did find interesting results. Regarding gender, I found that women are more likely to pursue health occupations and women are more likely to want to go far in college and academic reputation is important to them when choosing a college. Regarding race, I found that Asians are more likely to pursue STEM careers while African Americans are less likely to pursue STEM careers.

The next steps for this research would be to create models with more balanced data and a larger sample size. Additionally, the next HSLS survey is scheduled for 2024. It may be worth using questions from this survey when the results are released, or using a new outcome variable from this survey. Whatever the next steps taken with this research are, I feel that this research is a good starting point for removing barriers related to pursuing a STEM career. From my literature review, I found that the people who feel like they fit into STEM are more likely to pursue a STEM career, while people who don't feel like they fit in will not (Milgrom-Elcott). Thus, it is important to ensure that STEM careers, education, and promotion are welcoming and no one is treated differently or discriminated against due to their gender or race.

Specifically, in my literature review I found that women have greater representation and career satisfaction in the healthcare industry (Berlin et al.), but African Americans have less representation in STEM industries (Lennon). If we can create more policies and benefits for underrepresented groups in STEM, and eliminate the socioeconomic factors holding African Americans back, then underrepresented groups may also have greater representation and career satisfaction similar to women in healthcare. It is important that we keep this in mind so we as a society can create a better culture for underrepresented groups in STEM.

Acknowledgments

Thank you to Bradley Yam for helping guide me through the research process, from initially finding the data to drawing conclusions.

References

Berlin, Gretchen, et al. "Women in the Healthcare Industry." *McKinsey & Company*, 7 June 2019, www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/women-in-the-healthcare-industry. Accessed 8 Oct. 2022.

Frank, Kristyn. *A Gender Analysis of the Occupational Pathways of STEM Graduates in Canada*. 16 Sept. 2019. *Statistics Canada*, www150.statcan.gc.ca/n1/pub/11f0019m/11f0019m2019017-eng.htm. Accessed 8 Oct. 2022.

Lennon, Annie. "Why Are There So Few Black People in STEM?" *Labroots*, 14 June 2020, www.labroots.com/trending/chemistry-and-physics/17877/black-people-stem. Accessed 8 Oct. 2022.

Li, Russell, et al. "'Pushed toward STEM.'" *Best of SNO*, 10 Feb. 2022, bestofsno.com/55323/features/pushed-toward-stem/. Accessed 8 Oct. 2022.

Mau, Wei-Cheng J., and Jiaqi Li. *Factors Influencing STEM Career Aspirations of Underrepresented High School Students*. 2018. *Wichita University*,

soar.wichita.edu/bitstream/handle/10057/16274/Mau_Li_2018.pdf?sequence=1&isAllowed=y. Accessed 8 Oct. 2022.

McGee, Vanesha. "Guide to Diversity and Inclusion in STEM." Edited by Angelique Geehan. *ComputerScience.org*, 16 Sept. 2022, www.computerscience.org/resources/diversity-inclusion-in-stem/. Accessed 8 Oct. 2022.

Milgrom-Elcott, Talia. "To Pursue And Succeed In STEM, Students Need To Know They Belong." *Forbes*, 8 June 2022, www.forbes.com/sites/taliamilgromelcott/2022/06/08/to-pursue-and-succeed-in-stem-students-need-to-know-they-belong/?sh=149998643af4. Accessed 8 Oct. 2022.