Analysis of class surveys

Lucas Childs, Minu Pabbathi, Nathan Kim, Anna Liang, Bahaar Ahuja

2025-10-19

Executive summary

Our project focused on understanding how domain interests and the number of upper division courses taken relates to comfort in math, statistics, and programming. First, we looked at whether domain interest plays a role in comfort levels. After running a Kruskal-Wallis test, no statistically significant difference in comfort levels was found among domains. We then looked at whether diversity of interests affects comfort levels using an ordered logistic regression model and a decision tree. No predictive relationships were found between the number of interested domains and comfort levels. Finally, we examined the relationship between upper division coursework and comfort through k-means clustering, seeking to group these features. We found that students who have taken more courses tended to be grouped more with higher comfort levels, indicating that more coursework could improve comfort with math, statistics, and programming. Ultimately, while domain interest has little to no relationship with comfort, more coursework is associated with higher comfort.

Data description

Data were obtained by Google Forms survey that was administered to all students in PSTAT 197A during the Fall of 2025. The data was then split into two tables: background_clean and interest_clean. For our project, we combined the two tables and one hot encoded the individual domains. The variables in the final table are as follows:

Data Description

Variable	Description
response_id	Unique IDs to order observations (students).
prog.prof	Self-reported proficiency in programming (beg, int, adv).

Variable	Description
prog.comf	Self-reported comfort in programming (1–5).
math.prof	Self-reported proficiency in math (beg, int, adv).
math.comf	Self-reported comfort in math (1–5).
stat.prof	Self-reported proficiency in statistics (beg, int, adv).
stat.comf	Self-reported comfort in statistics (1–5).
updv.num	Number of upper-division classes taken (0–2, 3–5, 6–8, 9+).
dom_tf	Whether the student has domain specialization (Yes, No, Unsure).
rsrch_or_ind	Preference for research or industry project (industry, research, no preference).
Course columns (boolean)	Whether a student has taken each course. Columns: PSTAT100, PSTAT115, PSTAT120, PSTAT122, PSTAT126, PSTAT131, PSTAT160, PSTAT174, CS9, CS16, LING104, LING110, LING111, CS130, CS165, ECON145, PSTAT127, PSTAT134, CS5.
Domain interest columns (boolean)	Whether the student is interested in each domain. Columns: Biology, Chemistry, Ecology, Economics / Accounting, Entertainment, Environmental science, Music & Audio, Neuroscience, Psychology, Public health, Social or political science, Software development, Technology, media/musical technology.

Summary statistics for background-clean.csv (used for clustering analysis):

```
Summary Statistics:
       updv_numeric  prog.comf  math.comf  stat.comf
count     51.000000  51.000000  51.000000  51.000000
mean       7.882353   3.862745   4.039216   4.039216
std        2.635504   0.748855   0.773583   0.799019
min        1.000000   2.000000   3.000000   2.000000
25%        7.000000   3.000000   3.000000   3.500000
50%       10.000000   4.000000   4.000000   4.000000
75%       10.000000   4.000000   5.000000   5.000000
max       10.000000   5.000000   5.000000   5.000000
```

Feature correlations:

```
Feature Correlations:
Upper Div Courses ↔ Programming Comfort: 0.154
Upper Div Courses ↔ Math Comfort: 0.277
Upper Div Courses ↔ Statistics Comfort: 0.382
Programming Comfort ↔ Math Comfort: 0.217
Programming Comfort ↔ Statistics Comfort: 0.243
Math Comfort ↔ Statistics Comfort: 0.580
```

Questions of interest

We sought to understand what factors affect comfort levels in math, statistics, and programming. Our analysis addressed three different questions:

1. Does domain interest relate to comfort levels?
2. In addition, does the number of interested domains relate to comfort levels?
3. Can we group students based on the number of upper division courses they've taken and their comfortability in math, stats, and programming?

Findings

When examining the relationship between domain interests and comfort levels, we found that those most comfortable with programming tend to be interested in ecology, biology, economics/accounting. Comfort with math is most associated with interest in economics/accounting, music & audio, and public health. Finally, those most comfortable with statistics are interested in music & audio, biology, and public health.

No statistically significant differences were found in comfort across domains, implying that there is no connection between domain interest and comfort in math, programming, and stats. This could be because students in the capstone program are already comfortable in these fields and likely apply because they have experience in them.

Exploring the second question, we used an ordinal logistic regression model to test the null hypothesis that the number of domains a student is interested in has no association with their comfort (1–5) in programming, math, or statistics.

	outcome	n	odds_ratio	lower	upper	p_value	q_value
0	prog.comf	49	0.8878	0.6186	1.2743	0.5188	0.7781
1	math.comf	49	0.9822	0.7025	1.3733	0.9164	0.9164
2	stat.comf	49	0.8989	0.6509	1.2414	0.5176	0.7781

The odds ratios for all three subjects were near 1 (programming/statistics slightly < 1), all 95% confidence intervals included 1, and all q values (FDR-adjusted (Benjamini–Hochberg) p-values) were > 0.05. Thus, we found no significant evidence of the alternative and we did not detect a statistically significant association between the number of interested domains count and comfort levels in any subject.

To explore whether the number of domains a student is interested in predicts their comfort in programming, math, and statistics, I created a Domain Diversity Score (DDS) by counting unique domains per student and computed a Composite Comfort Index (CCI) as the average of their comfort ratings across the three subjects. I then trained a Decision Tree Regressor to model the relationship between DDS and CCI. While the model achieved a moderate R^2 of 0.278 on the training set, the test R^2 was -0.394, indicating poor generalization and performance worse than predicting the mean. The RMSE on the test set was approximately 0.72 comfort units. A visualization of actual versus predicted values showed a largely flat prediction line, suggesting little to no predictive power. Overall, these results indicate that the number of domains a student is interested in does not meaningfully predict their comfort level in programming, math, or statistics.

These results are consistent with Question 1, which also found no statistically significant link between domain interest and comfort in programming, math, or statistics.

To answer our question about grouping students by the number of upper division courses they've taken and their comfortability in math, statistics, and programming:

Using 5 cluster centroids, we grouped students who'd taken 6+ upper division courses into 2 groups of comfortability between 3.5 and 4 and one above 4. Then for students who'd taken 3-5 upper division courses, we discovered 2 groups of comfortability between 3 and 3.5 and the other above 4. Thus, on average, the clustering results show that we can group students who've taken more upper division courses into a category of relatively higher comfortability in math, stats, and programming. This intuitively makes sense, because as students complete more upper division coursework in math, statistics, and programming they would tend to feel more comfortable with those subjects. It's important to note that the clusters we've found do not imply causation, and that the silhouette score of the clustering, 0.29, was relatively close to 0, not indicating the strongest clustering configuration. Furthermore, correlations between the number of upper division courses taken and comfortability were relatively low (the highest being 0.38 between updv.num and statistics comfort).

```
Students with 9+ courses: 27
Average Programming Comfort: 4.00
Average Math Comfort: 4.26
Average Statistics Comfort: 4.26

Students with <9 courses: 24
Average Programming Comfort: 3.71
Average Math Comfort: 3.79
Average Statistics Comfort: 3.79
```

Conclusion

Overall, our analysis found that domain interest alone does not significantly relate to students' comfort levels in programming, math, or statistics. Both the Kruskal–Wallis test and the ordinal logistic regression provided no evidence of meaningful differences or associations. Additionally, the decision tree model using domain diversity scores showed poor predictive performance, further supporting the conclusion that the breadth of students' interests does not strongly influence comfort.

However, we did observe a clearer pattern when examining upper-division coursework: students who had completed more upper-division classes tended to cluster in groups with higher comfort levels across all three subjects. Although these results are not causal, they suggest that taking more advanced coursework is associated with increased comfort in technical domains, which aligns with intuitive expectations about skill development over time.

Limitations

Our analysis is based on self-reported survey data from a single class cohort, which may introduce biases in comfort ratings and domain interest reporting. The sample size is also limited, which may reduce statistical power and make it harder to detect smaller effects. Additionally, our models were intentionally kept simple (e.g., shallow decision trees, basic clustering), so more complex modeling techniques might uncover additional nuances.

Next Steps

Future analyses could explore these relationships more deeply by:

1. Collecting longitudinal data to examine how comfort evolves over time with coursework.
2. Expanding the sample size to include students from multiple cohorts or departments to increase generalizability.
3. Exploring interaction effects between domain interest and coursework, for example, whether domain alignment with coursework affects comfort differently.
4. Testing more advanced predictive models (e.g., random forests or gradient boosting) to assess whether nonlinear or interaction effects improve predictive power.
5. Incorporating qualitative data, such as interviews or open-ended survey responses, to better understand why some students feel more or less comfortable in these subjects.