STA304 Assignment 1

Clarence Chau Solai Ramu

Introduction

For high school students navigating the complicated pathways of course planning, university admissions, and scholarship opportunities; accessing high quality information and guidance is crucial. Yet research shows that for many students, especially those from underrepresented, first generation, or low income backgrounds, they face significant obstacles and challenges in accessing timely and relevant educational resources, guidance, and mentoring. For example, students without access to sufficient guidance counseling are more likely to be placed in less rigorous tracks and take fewer advanced courses, which can severely limit postsecondary options (Gándara & Bial, 2000). In addition, students from historically marginalized groups often rely more heavily on school based supports, since they may lack access to external networks or inside knowledge about university systems and scholarship processes (Quint et al., 2024).

The **STEMBuddies High School Handbook** was developed to address these challenges by providing a centralized, accessible resource that covers it all, for students in grades 9–12. However, as curriculum, admission requirements, scholarships, and student needs evolve, the handbook must be updated to remain relevant and responsive to students' experiences. The purpose of this survey is to gather direct feedback from high school students about which content areas of the handbook are most useful, which domains are underrepresented, and how the resource can better support their academic, personal, and emotional needs. By focusing on course planning, scholarship navigation, university admissions guidance, extracurricular opportunities, and mental health supports, areas often cited as critical to student success. We aim to discover gaps between the handbook's current design and real student priorities. The results will guide revisions to the handbook to ensure it serves as a current, inclusive, and effective tool for **all** students, especially those from underrepresented backgrounds in STEM.

Survey Showcasing

The survey for this project was designed and distributed through Google Forms and can be accessed at the following link: STEMBuddies Student Handbook Survey. It was developed

to assess high school students' perspectives on the STEMBuddies High School Handbook and to identify areas where additional support would be most impact. The survey included, an introduction outlining the purpose of the study, demographic questions to capture grade level and gender identity, needs assessment questions to evaluate priorities such as course planning and scholarship access, and a closing message thanking respondents and providing contact information. All questions are in the Appendix of this report for reference.

The demographic questions were designed to be inclusive but still concise. For example, the grade-level question used distinct categories ("Grade 9," "Grade 10," "Grade 11," "Grade 12") to allow straightforward responses. Gender identity was asked using multiple choice with an "Other (please specify)" option to ensure inclusivity while still facilitating quantitative analysis. Needs assessment questions employed Likert-scale responses (e.g., "Not Important (1)" to "Very Important (5)") so that students could communicate the relative importance of different resources. This format was chosen to allow both individual item level analysis and broader comparisons across content areas.

To highlight one key variable for analysis, the following question was selected from the needs assessment section:

Q5. How important is access to information about scholarships and financial aid in supporting your high school and postsecondary goals?
Response Options: Not Important, Slightly Important, Moderately Important, Important, Very Important

This question was chosen because financial support is the biggest decider of access to post-secondary education, particularly for students from more unfortunate situations. Its benefits include straightforward wording, alignment with the Likert scale structure used across the survey, and direct relevance to one of the central goals of the handbook. A potential drawback is that "scholarships and financial aid" are broad categories, which may encompass different programs or resources depending on the student's awareness. Narrowing this further could improve precision but would risk making the survey overly complex and time consuming. Thus, the broader phrasing was used to balance clarity and respondent burden.

Procedure

To realistically collect feedback on the STEMBuddies High School Handbook, the target population would be high school students in grades 9 through 12 across Canada. Because it would not be feasible to directly survey every student in this population, we propose a stratified convenience sampling procedure. Schools partnered with STEMBuddies and other outreach organizations could serve as the primary sampling frame. Within each school, student volunteers would be invited to complete the survey, with efforts made to ensure equal representation

across grades and genders. Each survey response would therefore represent an individual student as the sample unit. The strength of this approach is its practicality: it leverages existing networks while ensuring diverse grade level participation. However, potential biases may arise, including overrepresentation of students already engaged in STEM enrichment activities, as well as self selection bias if only the most motivated students choose to respond.

To simulate data for this project, we generated synthetic survey responses in R that mirror the structure and expected distributions of the original survey questions. For example, the showcased question on scholarship importance was simulated using a categorical variable with five Likert-scale levels: "Not Important," "Slightly Important," "Moderately Important," "Important," and "Very Important." A weighted probability distribution was applied to reflect realistic expectations (e.g., higher probabilities assigned to "Important" and "Very Important"), since financial aid is widely recognized as a priority among high school students. Similarly, demographic variables such as grade level were simulated with roughly equal sample sizes across grades to approximate a balanced design. Gender identity responses were simulated with multiple categories, including an "Other" option, to reflect inclusivity.

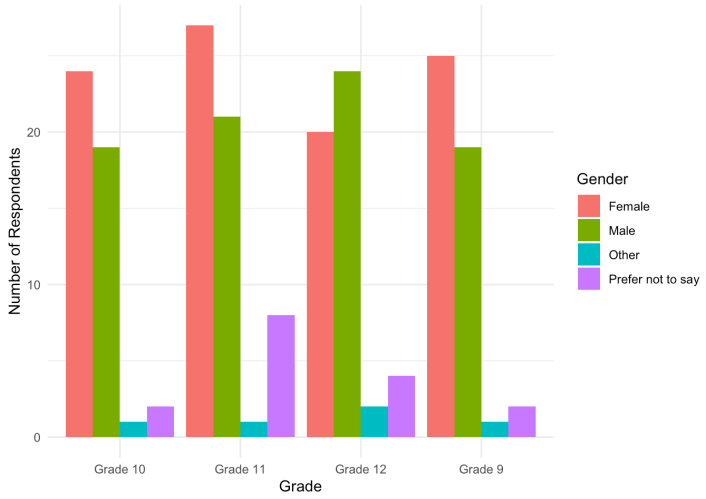
The simulation process was designed to be reproducible. For each question, we explicitly defined the response categories and sampling probabilities, and then drew responses using random sampling functions in R with a fixed random seed. This ensures that results can be replicated exactly in subsequent analyses. While not a perfect substitute for real-world data, the simulated dataset provides a consistent input for testing analysis methods and allows us to explore the potential relationships among variables.

Data

Our dataset was generated by simulating responses from 200 high school students across Grades 9–12. Each observation represents a unique student, with three sections of survey data: demographics, needs assessment of handbook content, and self identification with underrepresented groups in STEM. Data cleaning was minimal but deliberate: (1) categorical variables such as grade and gender were explicitly coded as factors with fixed levels to ensure consistent plotting, (2) multi select questions (e.g., underrepresented groups) were stored as binary indicator columns (1 = selected, 0 = not selected) with additional logic to handle "Prefer not to say" and skipped responses, and (3) Likert items were simulated as ordered factors ranging from 1 ("Not important at all") to 5 ("Extremely important"). By standardizing column names and ensuring every variable had clearly defined levels, we created a dataset that can be easily reproduced without relying on complex `tidyverse` functions.

The demographic variables serve primarily to contextualize the representativeness of the sample.

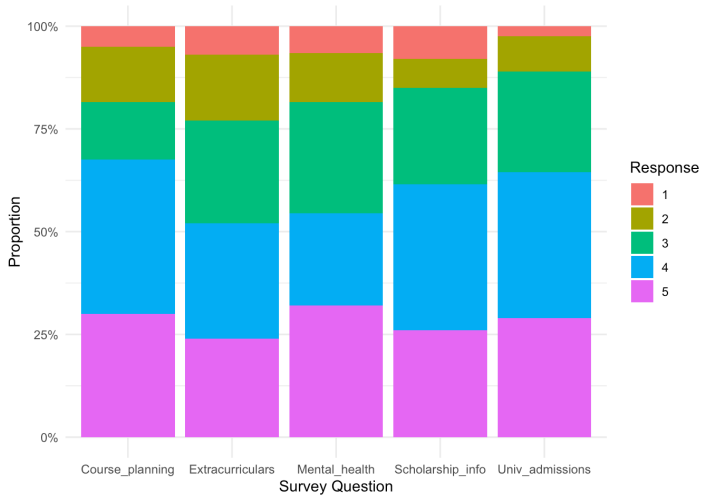
Figure 1: Demographic Breakdown by Grade and Gender



The sample was relatively balanced across grades, though Grade 11 showed slightly higher female participation. Both male and female respondents were well represented, with smaller but notable counts of students selecting "Other" or "Prefer not to say." This inclusion highlights gender diversity and ensures voices from all groups were captured. Overall, the demographic spread supports analysis across grade levels while acknowledging limitations in sample generalizability.

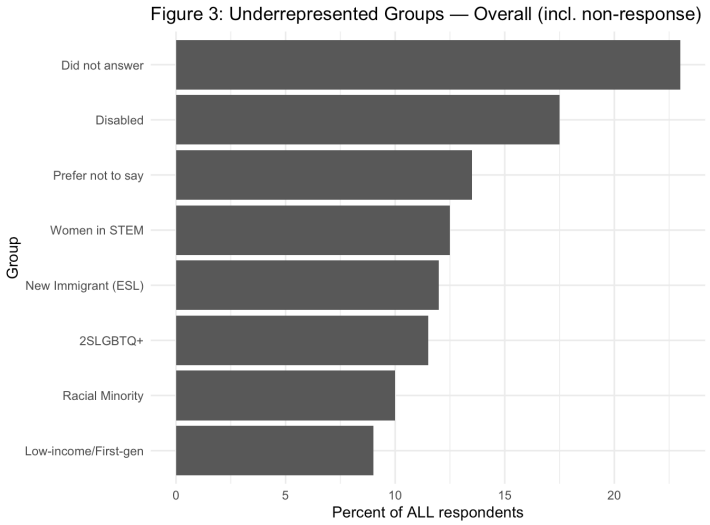
The main variables of interest are the five **needs assessment items**, which asked students to rate the importance of topics such as course planning, extracurriculars, mental health, scholarship information, and university admissions. Responses used a 5-point Likert scale, with higher values indicating greater importance.

Figure 2: Distribution of Needs Assessment Responses



Scholarship information, university admissions, and course planning were most frequently rated as "very" or "extremely" important (scores 4–5). Mental health and extracurriculars also received strong support but showed a slightly wider spread of responses, suggesting that while these areas matter, they may not be equally prioritized by all students. These patterns align with the survey's purpose, which is identifying handbook topics that require emphasis to best support underserved students.

The survey also included a multi-select question on **underrepresented group membership**. Students could identify with one or more categories (e.g., women in STEM, 2SLGBTQ+, new immigrant/ESL, racial minority, disabled, low-income/first-generation).



Roughly 20% of students did not provide a response, which may indicate discomfort with disclosure or survey fatigue. Or simply, that they are not a part of any underrepresented group mentioned. Among those who answered, meaningful representation was observed across categories, with particularly high counts in "Disabled" and "Women in STEM." This provides evidence that STEMBuddies is reaching diverse subgroups.

Finally, the data section sets up our inferential analysis. Later, we will calculate a confidence interval for one of the needs assessment variables. For example, the proportion of students rating **scholarship information** as "extremely important." From Figure 2, this item appears to have a notably high endorsement rate, making it both substantively interesting and statistically appropriate for interval estimation. Establishing a margin of error around this proportion will help quantify how confident we can be that high scholarship information demand extends beyond our sample.

Methods

Our primary inferential analysis focuses on the proportion of students who rated *scholarship information* as "extremely important" in the needs assessment section of the survey. Because this outcome is categorical (students either did or did not select the highest rating), the appropriate statistical tool is a **confidence interval for a population proportion**. This allows us to estimate, with a level of confidence, the true proportion of all students who would give this response if the survey were administered to the broader population.

The general formula for a confidence interval for a proportion is:

$$\hat{p}; \pm; z\_\alpha/2 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Here, $\hat{p}$ is the sample proportion of students in our dataset who selected "extremely important," n is the total sample size, and $z\_\alpha/2$ is the critical value from the standard normal distribution corresponding to the chosen confidence level (for a 95% confidence interval, $z\_\alpha/2 \approx 1.96$. The square root term is the **standard error**, which measures how much the sample proportion is expected to vary due to random sampling.

This approach assumes that the sample size is large enough for the sampling distribution of $\hat{p}$ to be approximately normal, a condition met here with 200 respondents. Interpreting a 95% confidence interval means that, if we repeated the study many times, about 95% of the intervals constructed in this way would contain the true population proportion. This method therefore measures and quantifies the uncertainty around our estimate of how important scholarship information is to students, preparing the reader to interpret the numerical results in the next section.

Results

We estimated the proportion of students who rated *scholarship information* as "extremely important" and constructed a 95% confidence interval for this estimate. The results are shown in Table 1.

Table 1. Estimated proportion and 95% confidence interval for students rating scholarship information as extremely important (n = 200).

Estimate	95% CI Lower	95% CI Upper	N
0.26	0.199	0.321	200

Table 1 indicates that approximately 26% of surveyed students rated scholarship information as most important. The 95% confidence interval ranges from about 20% to 32%, suggesting that even when accounting for sampling variability, only about one-quarter to one-third of the student population would be expected to prioritize scholarship information at the highest level.

These findings are consistent with earlier descriptive analyses, which showed strong but not universal endorsement of financial aid resources. The interval width is relatively narrow, reflecting the stability of the estimate with a sample size of 200. This result appears reasonable given the context, while scholarships are a significant concern for many students, not all view them as the single most critical element compared to other academic or well-being needs.

Taken together, these results highlight a clear but selective demand for scholarship guidance, which directly informs STEMBuddies' efforts to balance financial, academic, and personal support in its handbook. The following Discussion section expands on these implications and considers how they align with the broader mission of supporting underserved students in STEM.

Generative AI / Workflow Statement

For this assignment, we used generative AI tools (specifically ChatGPT) to support my workflow in several ways:

- **Coding Support in R:** We used AI to help debug portions of the R code for simulating survey data (e.g., generating Likert-scale responses, handling multi-select variables, and producing plots such as histograms and bar charts).

- **Idea Generation:** We consulted AI for brainstorming survey questions aligned with the STEMBuddies handbook goals, and for formatting possible report structures.
- **Formatting:** We consulted AI to help us format the document in ways to fit the assignment criteria, and to ensure a neat and clarity rich report.

To ensure that the final report was our own work, we carefully reviewed and revised all AI generated content. We verified citations independently to confirm their relevance and accuracy, and we rewrote text where needed to align with our own understanding and the assignment requirements. For R code, we ran and tested the code ourselves in RStudio, debugging and modifying until the output matched the expected structure.

In summary, AI tools were used to *supplement* our formatting, coding guidance, and brainstorming ideas, but the final analysis, critical thinking, and interpretations are our own.

Bibliography

American Statistical Association. (2016). *Guidelines for survey research methods.* Alexandria, VA: ASA.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Thousand Oaks, CA: Sage.

Museus, S. D., Palmer, R. T., Davis, R. J., & Maramba, D. C. (2011). *Racial and ethnic minority students' success in STEM education.* ASHE Higher Education Report, 36(6), 1–140.

National Academies of Sciences, Engineering, and Medicine. (2016). *Barriers and opportunities for 2-year and 4-year STEM degrees: Systemic change to support students' diverse pathways.* Washington, DC: National Academies Press. https://doi.org/10.17226/21739

STEMBuddies. (2025). *STEMBuddies high school handbook.* Retrieved from https://stembuddies.ca

Appendix

A. Survey Questions

Introduction Message:
Thank you for participating in the STEMBuddies High School Handbook Survey. Your responses will help us improve the handbook so it better supports students like you in areas such as course planning, scholarships, admissions, extracurriculars, and mental health. All responses are anonymous and will only be used for research purposes.

Section 1: Demographic Questions

1. **What grade are you currently in?**
    - Grade 9
    - Grade 10
    - Grade 11
    - Grade 12
2. **What is your gender identity?**

- Male
- Female
- Other (please specify)
- Prefer not to say

3. **Do you identify as belonging to one or more of the following groups underrepresented in STEM?** (Select all that apply) (optional)
   - Women in STEM
   - 2SLGBTQ+
   - New immigrant / ESL
   - Racial minority
   - Disabled
   - Low-income / first-generation student
   - Prefer not to say

Section 2: Needs Assessment

4. **How important is access to information about course planning in supporting your academic goals?**
   - Not Important
   - Slightly Important
   - Moderately Important
   - Important
   - Very Important
5. **How important is access to information about scholarships and financial aid in supporting your high school and postsecondary goals?**

- Not Important
- Slightly Important
- Moderately Important
- Important
- Very Important

6. **How important is access to information about university admissions requirements and processes?**
   - Not Important
   - Slightly Important
   - Moderately Important
   - Important
   - Very Important
7. **How important is access to information about extracurricular opportunities?**
   - Not Important
   - Slightly Important
   - Moderately Important
   - Important
   - Very Important
8. **How important is access to resources and information about mental health?**
   - Not Important
   - Slightly Important
   - Moderately Important
   - Important
   - Very Important

Closing Message:
Thank you for completing the survey! Your feedback will directly inform updates to the STEM-Buddies High School Handbook. If you'd like to learn more about our work, please visit our website.

B. Glimpse of Simulated Data

```
library(knitr)
library(kableExtra)

survey_data <- readRDS("survey_data.rds")
df <- head(survey_data, 10)

# Optional: nicer headers
colnames(df) <- gsub("_", " ", colnames(df))

tab <- knitr::kable(
  df,
  format   = "latex",
  booktabs = TRUE,
  longtable = FALSE,  # keep on one page so rotation is clean
  escape   = FALSE,
  caption  = "Appendix Table B1: First 10 rows of simulated survey data."
) |>
  kable_styling(
    latex_table_env = "tabularx",  # enables X/Y columns (auto-wrap)
    full_width = TRUE,
    font_size = 8,
    position = "center"
  ) |>
  # Fix first two columns to readable widths; wrap the rest with 'Y'
  column_spec(1, width = "8em") |>
  column_spec(2, width = "8em") |>
  column_spec(3:ncol(df), latex_column_spec = "Y")

landscape(tab)
```

Table 3: Appendix Table B1: First 10 rows of simulated survey data.

Grade	Gender	Scholarship info	Course planning	Univ admissions	Extracurriculars	Mental health	Underrep Not-Answered	Women in STEM	Racial Minority	2SLGBTQ+	Low-income/First-gen	New immigrant (ESL)	Disabled	Prefer not to say
Grade 11	Female	4	3	3	4	5	1	NA	NA	NA	NA	NA	NA	NA
Grade 9	Prefer not to say	4	3	5	5	4	0	0	0	0	0	0	0	1
Grade 11	Male	4	3	4	3	5	1	NA	NA	NA	NA	NA	NA	NA
Grade 9	Male	5	4	2	2	5	0	0	0	1	0	0	1	0
Grade 9	Female	5	4	3	4	3	0	0	0	1	0	0	1	0
Grade 10	Male	5	5	4	4	3	0	1	1	0	1	0	0	0
Grade 12	Female	5	4	4	5	4	0	0	0	0	0	0	0	1
Grade 9	Female	4	3	3	4	2	0	1	0	0	0	1	0	0
Grade 12	Female	4	5	3	5	4	0	0	0	0	0	0	0	1
Grade 11	Female	4	4	5	3	1	0	0	0	0	0	1	0	0