Bias Mitigation, Reporting and Analysis: Surveying High School Educators

Lucas Anthony

Aroona Atmaram

Ruchi Kumar

Professional Studies, Northwestern University

MSDS 402: Introduction to Data Science

Dr. John Derwent

June 7, 2020

Introduction

The next steps in designing a successful survey are bias identification and mitigation, an analysis plan for the results, and a process for reporting survey response progress. These steps will ensure the results of the survey are credible and meaningful to the stakeholders, so they can make informed and impactful decisions. Completing these activities in the planning stages will allow for adjustments to be made prior to launching the survey.

Bias Consideration and Mitigation Plan

We have identified several key areas where bias could be introduced in our survey that we will discuss in detail. This survey is intended to gain a better understanding of what subjects educators find most conducive to online instruction. The primary mode of delivery that was chosen for this survey was online, via email request. Bias may be introduced due to the fact that an online mode is chosen to ask educators about their impressions of online education. Educators that are less tech savvy may have limitations in responding to the survey. More importantly, educators that have issues with internet access will be limited in their ability to respond. This is particularly important as internet access is a key area of interest we are seeking to explore. In order to combat this potential for bias, we are including the opportunity for respondents to opt into a phone survey. This is to allow those that are less comfortable with computers, or have internet access issues, to have the opportunity to take part in this survey.

Another area for bias is in potential misrepresentation of the population. While we have chosen to stratify by region, the survey is asking respondents about the county they teach in. The surveyors have classified each county in New Jersey into one of three regions so that this data can be tracked as results come in. There is potential for more heavily populated regions to become overrepresented and this needs to be addressed. The same bias opportunity comes into play with the sample responses of educators based on subject taught. This is another key demographic that will be tracked as results come. If regions, or subjects, are not adequately represented, we will push additional reminders to those underrepresented regions. If there is a continued bias of overrepresentation, the surveyors will add weight to the responses, which will be discussed in more detail below.

We have some additional concern for the potential of answer bias. While the questions in the survey may not be considered sensitive, the surveyors are sensitive that the survey is being sent through the New Jersey Educational Association to ask employees about their thoughts on online learning. While we state several times that the survey is completely anonymous, there can still be mistrust from employees in believing their responses are completely anonymous. In order to address this concern by respondents, we have opted to begin the survey with the ask for the respondents email address if they wish to enter the sweepstakes. We will also be asking the participants to enter the email address they wish to use for the sweepstakes, providing another opportunity for anonymity. We will reiterate the anonymity of the responses and respondents will be able to see immediately that the question is completely optional and is intended to alleviate concerns early.

Analysis Plan_

Our analysis plan consists of coding, missing value imputation, weighting, exploratory data analysis and multinomial logistic regression.

Coding. Our survey has 11 multiple choice questions.The text responses will be coded as integers to facilitate analysis.The direction of scale is kept consistent for all the questions; the low end of the scale is negative responses, with the top of the scale reserved for the most positive responses. Highest score for positive response and lowest score for negative response.

Data Quality. The impact of poor data quality is the potential of creating bias and the opportunity for misleading results. To ensure data quality, we will take measures to provide respondents with a good user experience. The web survey will be programmed to adapt to both PC browsers and mobile browsers. We would provide tooltips with examples and scenarios that will help them answer questions. The questions are grouped to follow a pattern to help the respondent navigate the survey. Though these measures will be taken, it is assumed there will still be missing data and outliers, which will require cleansing.

Data Cleansing. As this survey is web based, the key questions have been made mandatory to avoid missing values. Key questions are those that are particularly asking about the subjects and experience with teaching those subjects. To address the missing value of non-mandatory questions, excluding open-ended questions, the following method has been used: Mode value from the respective strata will be used to impute the missing value. If there are greater than 30% missing values we would choose to drop the question for analysis.

Open-ended responses will be cleaned and categorized to further analysis. Data will be broken down into text units and then sorted into themes or categories using qualitative analysis tools such as HyperRESEARCH or NVivo. These tools allow for the creation of a coding scheme that can apply to qualitative data (Wilson, C. 2013).

Exploratory Data Analysis. Our survey research chooses exploratory data analysis for visual summaries of the response data. We start by plotting the response to determine the descriptive statistics of the variables like skewness, kurtosis, median and mode. Graphs such as box plot, histogram, scatter plot and bar charts will be used when appropriate. The response for the question on the primary subject taught will be plotted against other responses to determine correlation and draw patterns. We will be categorizing by subject and plotting educator's responses on questions pertaining to conduciveness of online learning against internet and technology accessibility. We will also review outliers for trends before dropping them from the analysis to avoid skewing the results. The bag of words technique will also be applied and infographics created with most used words for the open-ended responses.

To further draw inference from the data, we plan to fit a Multinomial Logistic Regression model and make a prediction. Because the key questions in the survey are categorical rather than binary, this model was selected. We would choose the primary subject as the multinomial dependent variable and use the different responses as independent variables to draw a prediction.

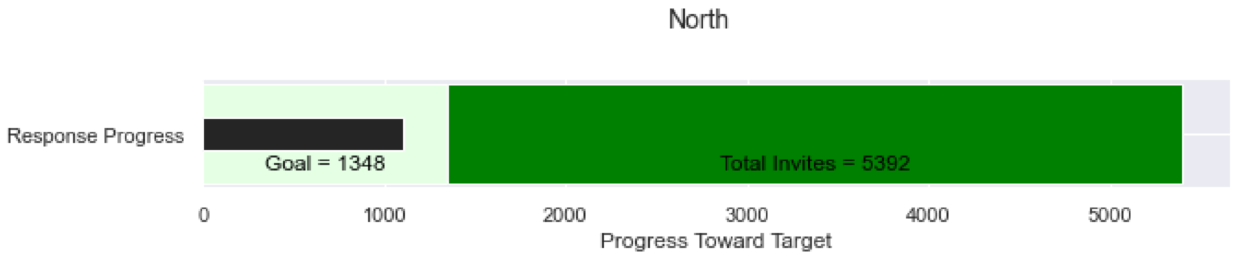
Weighting. We plan to use weighting to adjust survey statistical computations to account for non-response. If a particular region/subject is underrepresented, weighting will be applied to account for non-response in that region/subject strata.

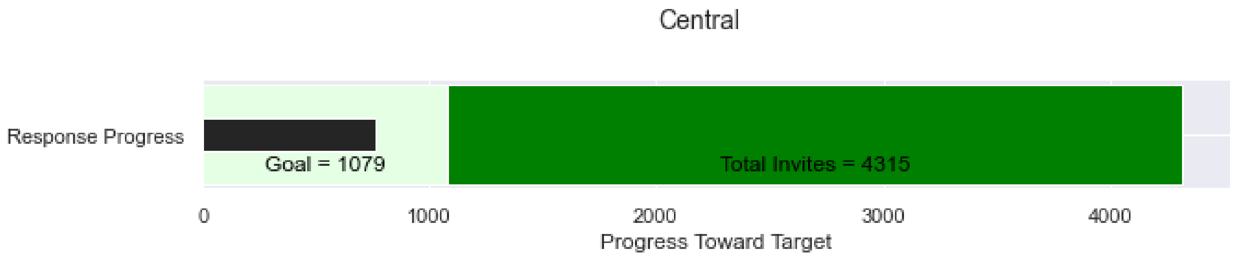
Reporting Plan

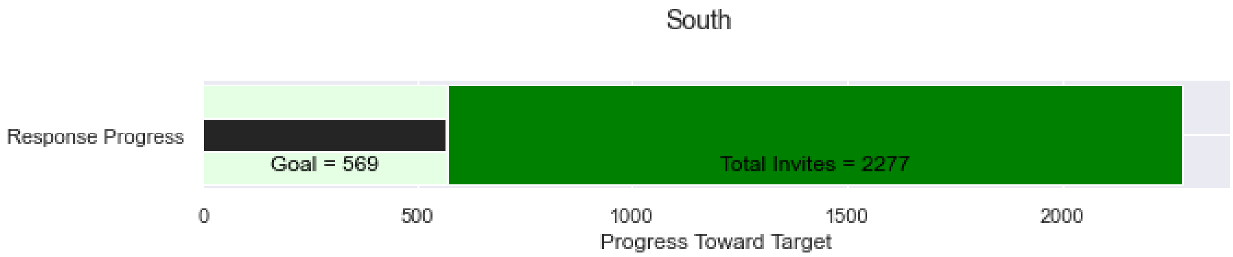
A weekly report will be sent out to keep the team apprised of what progress is being made in securing responses to the survey. This report will include the target number of responses, number of surveys started, number of surveys completed in the current week, and the total number of surveys completed to date. Each of these will be reported by strata. The report will also have the start date of the survey, the projected end date, and the current date.

The objective of this report is to understand progress towards completing the survey within the planned time. By knowing which strata may be under-represented, we can decide how to proceed: attempting to increase the response rate with increased contact, or consider over-sampling that strata.

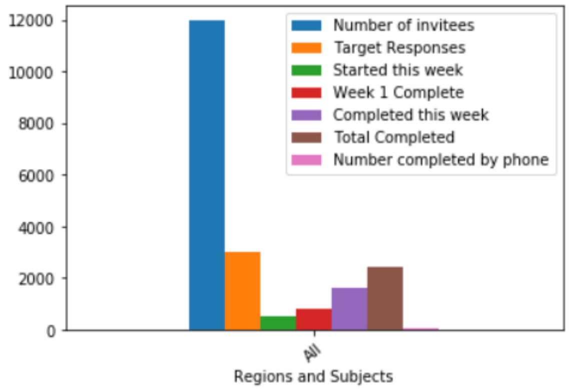
Week 2 Report

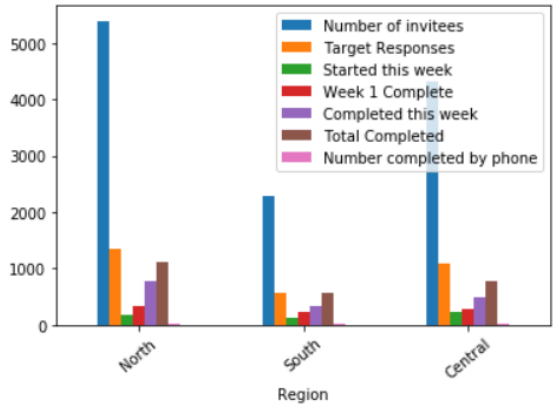


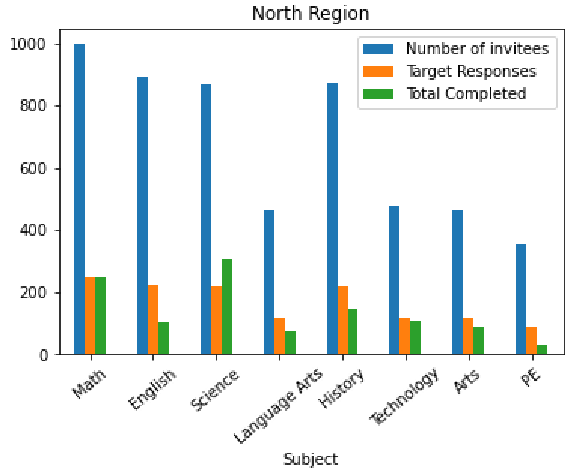


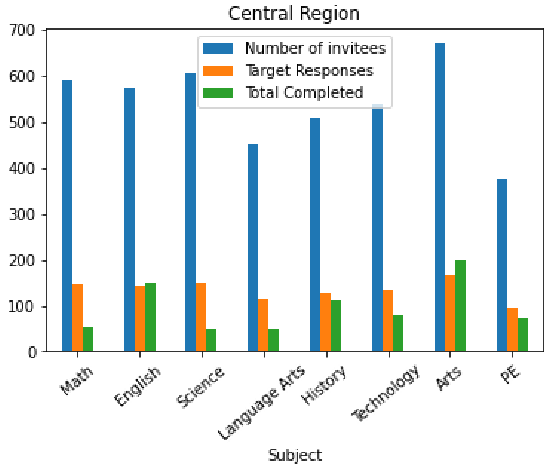


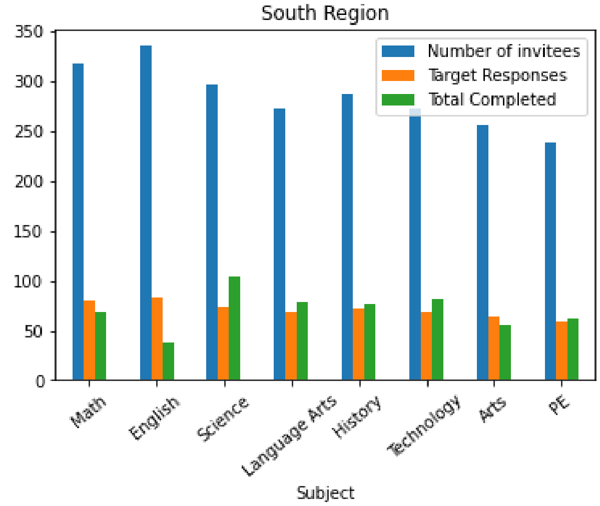
Total Progress











Conclusion

This survey may be susceptible to mode bias, by surveying for experience in teaching online via the web, population bias if certain strata are underrepresented, or answer bias for questions that are perceived as sensitive. We are taking efforts to mitigate all of these through offering phone interviews, watching progress of respondents in each strata and highlighting anonymity of the survey. The results of this survey will be analyzed using exploratory data analysis and multinomial logistic regression, after coding and any necessary imputation and weighting have taken place. To keep the team apprised of progress on data collection, a weekly report will be created and circulated to track responses.

References

Wilson, C. (2013). *Credible checklists and quality questionnaires: A user-centered design method*. Waltham, MA: Morgan Kaufmann.

Appendix:

Survey Start: 6/15/2020 Survey End: 7/5/ 2020 Current Week: 6/29/2020

Region	Subject	Target Responses	Surveys Started	Complete this week	Total Completed	% Complete by phone
All	All Subjects	2996	519	3428	5314	82
North	Math	249	25	305	388	4
	English	223	14	200	244	9
	Science	217	28	306	506	0
	Language Arts	116	53	85	155	2
	History	219	32	206	286	4
	Technology	119	14	147	247	3
	Arts	117	6	90	170	5
	PE	89	2	60	80	4
Central	Math	148	7	140	162	7
	English	143	14	201	290	7
	Science	152	9	130	250	3
	Language Arts	113	22	36	60	1
	History	128	43	189	310	0
	Technology	135	35	223	379	0

	Arts	167	55	218	399	6
	PE	94	32	97	163	2
South	Math	80	40	148	188	5
	English	84	5	68	98	5
	Science	74	23	147	245	0
	Language Arts	68	21	53	89	0
	History	72	15	130	216	6
	Technology	68	10	76	141	4
	Arts	64	9	85	105	2
	PE	60	5	88	143	3