Evaluating Student Perception of Assessment in Introductory Quantitative Studies

STEM occupations are the fastest growing jobs in the workforce; therefore, the demand for university graduates with STEM degrees is at an all-time high. Ideally, students do not only feel motivated to pursue STEM for the sake of job opportunities, but also to learn the material. Previous research shows that student perception influences student motivation, but does not discuss student perception of assessment, a key measure of academic achievement. In this thesis, we look at the intersection of student perception and assessment in STEM. We define student perception of assessment as how adequately a student believes an assessment measures his/her learning. We aim to quantify relationships between student perception of assessment and student identity, class format, and grading in introductory quantitative studies courses, a sector within STEM education. We survey 757 students from five different introductory courses. We find that the odds a student believes an assessment measures their learning are higher for projects than exams, for students who predict they will receive higher grades, and for students who major in quantitative courses than other subjects. Research on students' perception of assessment in STEM has the potential to restructure how instructors teach, assess, and grade students at the university level.

Introduction

Science, technology, engineering, and math (STEM) occupations are growing at the fastest rate out of any sector in the workforce. The U.S. Bureau of Labor Statistics shows employment in STEM occupations has grown 79 percent in the past three decades and STEM jobs are projected to grow 11 percent from 2020 to 2030 (U.S. Bureau Labor Statistics 2022). This demand stems from the widespread understanding that STEM is needed to meet the challenges the United States faces in "energy, health, the environment, and national security" (Lynch, Peters-Burton, and Ford 2014). In recent decades, "there has been an explosion in interest, investment, programs, research, and data" towards finding a solution to meeting the demand for STEM workers (White and Shakibnia 2019). However, "over 90 percent of all STEM occupation require at least some post-secondary education," (White and Shakibnia 2019). As a result, there is a serious need for students with a university-level understanding of STEM topics. Therefore, universities play a critical role in getting students through STEM pipeline to meet this demand for workers. We see that universities are more motivated to better prepare their students in STEM than ever before. Around the world, there have been major reforms to improve STEM education at all levels. In the U.S., in 2011, former President Barack Obama announced his goal of adding 100,000 STEM teachers to the nation's classrooms (Will 2022). It is evident that universities are pumping time, money, and resources into education reform in order to address the broader goal of preparing more students to work in STEM. We also see that more students are majoring in STEM than ever before. According to a report from the National Student Clearinghouse, the proportion of STEM degrees has increased dramatically between 2004 and 2014 at the bachelor's, master's, and doctoral levels (Bidwell 2015).

Although these investments in education and lucrative job opportunities may motivate students to pursue STEM degrees, they may not motivate students to learn STEM material. Previous research (Brown et al. 2016) shows that student motivation is influenced by student perception, but does not discuss student perception of assessment, a key measure of academic achievement in STEM courses. Thus, in this paper, we look at the intersection of student perception and assessment in STEM. We define student perception of assessment as how adequately a student believes an assessment measures his/her learning. We aim to quantify relationships between student perception of assessment and student identity, class format, and grading in introductory quantitative studies courses, a sector within STEM education, at SCHOOL University. Research on students' perception of assessment in STEM has the potential to restructure how instructors teach, assess, and grade students at the university level so that they learn best.

Data

Data Collection

In this study, we surveyed students in quantitative studies courses at SCHOOL University in an effort to answer our specific research question. We collected the data used for this thesis research in the Fall 2022 semester at SCHOOL University under IRB Protocol #[2022-0545]. The IRB was developed with consultation with SCHOOL Learning Innovation during Summer 2022. The data was managed by SCHOOL Learning Innovation.

We first sent emails to instructors teaching introductory statistics, computer science, and math courses at SCHOOL inviting them to participate in the study and obtaining their permission to send two surveys to their students and collect their grades (in grade ranges) on two assessments.

The assessments include exams and projects that make up a substantial percentage of students' final grades. We only sent surveys to the classes of instructors who consented to participate. We define introductory courses as those which do not have any prerequisites for students to enroll.

Both surveys were administered at the end of an assessment as a required part of the class. The first assessments were generally close to the mid-semester point and the second assessments were more cumulative assessments at the end of the semester. Depending on the course, surveys were either distributed via email, or during class. We also sent consent forms to all of the students in every course (Appendix A). The students in the study are SCHOOL undergraduates aged 18 and older. Students were asked to participate in this study by consenting to share their survey responses and grades from the associated courses for research purposes. Only the survey responses of students who consented were used in this research.

After course grades were posted, we asked each professor to send SCHOOL Learning Innovation their students' grades on the two selected assessments in grade ranges specified by the researchers. Professors were asked to share grades for all students, but the research study only used grades from students who consent to participate. Lastly, after receiving all survey responses, SCHOOL Learning Innovation sent anonymized survey responses to each professor from their respective courses. These responses served as sources of feedback for professors on their assessments and were part of the incentive for faculty to participate in the research.

Survey Instruments

The specific instruments for this project include two online surveys administered using Qualtrics. The goals of each survey are to first, learn some information about the individual students that may impact their perceptions and survey responses, and second, understand students' perceptions of their assessments as reflections of their learning. We surveyed students on two assessments instead of one for two main reasons: we wanted to capture two types of assessments (exams versus project); we wanted to capture assessments that are cumulative versus not. The first survey instrument can be seen in Table 1 and the second survey instrument can be seen in Table 2.

Table 1: First Survey Instrument

Question	Response
What is your year in school?	Select: Freshman, Sophomore, Junior, Senior
What your major or intended major?	Text response
Respond to the statement: This exam adequately measured my learning.	Select: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
Explain your response.	Text response
What grade do you think you will receive on this exam?	Select: Below 70, 70-75, 75-80, 80-85, 85-90, 90-95, 95-100

Table 2: Second Survey Instrument

Question	Response
What is your (intended) first major?	Text response

Question	Response
What your (intended) second major? If you do not (intend to) have a second major, please skip to the next question.	Text response
Respond to the statement: This exam adequately measured my learning.	Select: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
Explain your response.	Text response
What grade do you think you will receive on this exam?	Select: Below 70, 70-75, 75-80, 80-85, 85-90, 90-95, 95-100

In order to determine the questions asked in the surveys, we first looked at previous research to see whether there was a viable existing survey instrument. We found survey instruments including student evaluations of teaching, satisfaction, and course perceptions that influenced our own design. Ramsden and Entwistle's Course Perceptions Questionnaire (CPQ), which was developed to "measure the experiences of British students in particular degree programmes and departments most closely aligned with the goals of our survey instrument" (Richardson 2005). The CPQ, however, contained 40 items in eight scales that reflected different aspects of effective teaching. The CPQ and the other survey instruments mentioned above motivated our use of a Likert scale and grade scale in our survey, however we could not use these exact surveys for our research.

We intentionally chose to keep the surveys short with the hope that students would be more willing to fill out the surveys. This decision was based off of SCHOOL Learning Innovation's experience in previous projects. We asked students to explain their Likert scale responses so that we would have more context. Furthermore, we originally had intended to only ask students about their major in the first survey; however, we decided to add questions about major back into the second survey. We decided to restructure the data collection on major from a single open text box where students were instructed to input their first and second (intended) majors, to two text boxes, asking for first (intended) major and then separately for second (intended) major (if any), so that we could get clearer responses. In the first survey, we anticipated that students might not put multiple majors, or that students who put multiple majors might not clarify which major was first versus second.

Data Preparation

Data managers from SCHOOL Learning Innovation joined responses from surveys 1 and 2 and student grades from each class into a single data set, and then de-identified the data and removed students who did not provide consent for research. Due to privacy concerns, each student was provided with a unique ID, and the names and net IDs of the students were removed from the dataset. For the dataset utilized in this thesis, students can solely be identified by their row number. A separate dataset serves as a link between the individual students and their names and net IDs. The separate dataset can only be accessed by SCHOOL Learning Innovation, not the researchers. Students in multiple courses make up less than five percent of the total students so we treat them as independent and assign different IDs across different courses. Students who submitted multiple responses in the same class are assigned the same ID. SCHOOL Learning Innovation removed student ID from the dataset after the anonymization process.

The datasets were originally separated by class and organized as rows of students. For each class, we split the rows into student and assessment pairs. We categorized actual grades into buckets to

create the actual grade variable. Finally, after compiling the data, we dropped rows with missing data in columns excluding description because we deemed these responses to be incomplete. As a result, we dropped nine rows. After this initial data wrangling, there were 757 rows and 7 attributes per row. The first few rows of the dataset compiled in this project are available below.

Table 3: First few rows of dataset

course	year	major	measureLearn	description	gradePred	gradeActual	assessNum
A	Junior	Neuroscience	Agree	It challenged ...	90-95	90-95	1
A	Sophomore	Economics	Agree	I was ...	80-85	70-75	1
A	Sophomore	Economics	Agree	I thought ...	85-90	85-90	1
A	Sophomore	Undecided	Agree	NA	85-90	75-80	1
A	Junior	Computer Science	Agree	Nothing was ...	85-90	90-95	1
A	Sophomore	Sociology	Strongly Agree	We covered ...	90-95	85-90	1

The variables in the dataset are as follows:

Table 4: Data Dictionary

Name	Description
`course`	Name of the introductory statistics, math, or computer science course labeled as A, B, C, D, or E
`year`	Year in school (freshman, sophomore, junior, or senior)
`major`	Major or intended major in school
`measureLearn`	Response to the statement: "this assessment adequately measures my learning" on a Likert scale
`gradePred`	Students' prediction of their grade on assessment 1 in ranges: below 70, 70-75, 75-80. 80-85, 85-90, 90-95, 95-100
`gradeActual`	Students' actual grade on assessment 1 in ranges: below 70, 70-75, 75-80. 80-85, 85-90, 90-95, 95-100
`assessNum`	Encoding of the first assessment or second assessment

Response Rates

We asked students in five different courses to fill out two surveys in the Fall Semester of 2022 at SCHOOL University. We asked 889 students to participate in our research. Four-hundred and three students filled out both surveys and consented to participate in research. Our overall response rate was approximately 45%. The response rates within each courses varied substantially, with the highest response rate being 80% and the lowest response rate being 28%. On average, the response rate for each courses was approximately 52%.

Methods

Modeling

For our data, we use a logistic multilevel model to assess the relationship between student perception of assessment and student identity, class format, grading in introductory quantitative studies courses at SCHOOL University. In our research, we group students' responses to the statement "this

assessment adequately measured my learning" as either yes or no. Students respond to the survey question on a Likert scale, and we categorize "strongly disagree", "disagree," and "neutral" as "no" responses, and we categorize "strongly agree" and " agree" as "yes" responses. Therefore, our binary response variable indicates whether students believe an assessment reflects their learning or not. We decide against using ordinal logistic regression because we aim to investigate the differences between agreement and disagreement rather than the differences between the Likert scale responses, i.e. strong agreement and agreement. Since logistic regression is characterized by research questions with binary responses, this model is appropriate for our data. Moreover, we use a multilevel (two level) model. The response variables and covariates in our data have been collected at two, nested levels (Figure 2).

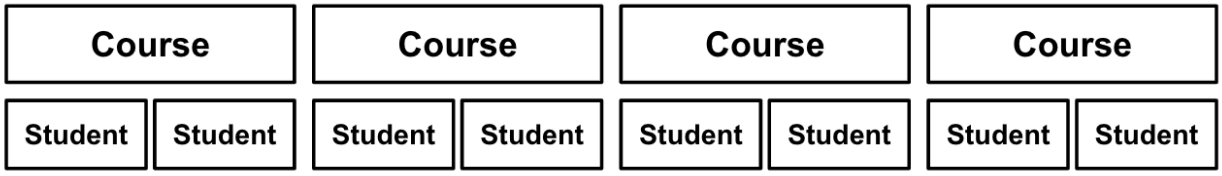

Figure 1: Nested Levels

Level two refers to the largest observational unit, the course. Level one refers to variables measured at the most frequently occurring observational unit, the student within each course. We use a multilevel model with this dataset because we cannot reasonably assume that responses are independent by course. We expect that students in a course with the same instructor, course size, and other course specific attributes may respond similarly. We account for the fact that students in the same course may share similar response patterns by treating it as a random effect. However, we do assume that responses are independent by student. We considered creating a three level model, with course as the third level, assessment number within course as the second level, and student within assessment and within course as the first level; however in the context of our research questions we are more interested in estimating and reporting values across assessment type rather than controlling for the time of the assessment. Since assessment type is not nested within each course, meaning some courses were surveyed on multiple assessments of the same type, this variable cannot be considered a level. Moreover, we cannot include assessment number and assessment type in the model since the variables explain essentially the same information. As a result, we include assessment type as a variable at the student level and exclude assessment number from the model. The model can be found below, where p_{ij} equals the probability of having selected agree or strongly agree to the statement: 'this assessment adequately measured my learning.'

Level One:

$$log(\frac{p_{ij}}{1-p_{ij}}) = a_i + b_{1i}x_{1ij} + ... + b_{zi}x_{zij} \quad \text{where } z = \text{number of predictors}$$

Level Two:

$$a_i = \alpha_0 + \tilde{u}_i \text{ where } u_i \sim N(0, \sigma_u^2)$$
$$b_{ki} = \beta_k \quad \forall k \in \{1, ..., z\}$$

Composite Model:

$$log(\frac{p_{ij}}{1-p_{ij}}) = \alpha_0 + \beta_1 x_{1ij} + ... + \beta_z x_{zij} + u_{ij}$$

We do not have any evidence to suggest that any assumptions are violated, that would prevent us from using a logistic multilevel model. A brief assessment of the assumptions can be found in Appendix B. All model analysis was conducted using the `lme4` package in R (Bates et al. 2015).

Text Analysis

In both surveys, the third question asked students to respond to the statement: "this exam/project adequately measured my learning" on a Likert scale. The fourth and second to last question on the surveys asked students to explain their responses in a free response text box in Qualtrics.

We analyze these text responses via sentiment analysis. There are a variety of methods and dictionaries that exist for evaluating the opinion or emotion in text. The `tidytext` package provides access to several sentiment lexicons (Silge and Robinson 2016). Three general-purpose lexicons we consider are AFINN from Finn Årup Nielsen, Bing from Bing Liu and collaborators, and NRC from Saif Mohammad and Peter Turney. AFINN and NRC lexicons tend to overestimate positive sentiment. In our research we do not want to overestimate positive sentiment and want to ensure that students truly perceive that assessments measure their learning. Thus, we use the Bing lexicon for the majority of our analysis. Furthermore, the Bing lexicon categorizes sentiment into positive, negative, and neutral, whereas the NRC lexicon splits into additional categories (anger, anticipation, fear, joy, surprise, etc.) which are not necessary given short length of text responses. To calculate the sentiment score, we use the following formula:

$$\frac{\#\text{positive words} - \#\text{negative words}}{\#\text{total words}}$$

It is important to note, however, when looking at bigrams, we use the AFINN lexicon. Here, the sentiment scores we calculate with the Bing lexicon do not accurately reflect the sentiment of pairs of words like the AFINN lexicon. The AFINN lexicon assigns words a sentiment value between -5 and 5 and therefore can more accurately reflect the true score of a pair of words. Therefore, we use the AFINN lexicon when handling bigrams and the Bing lexicon otherwise. To calculate the sentiment score of the bigrams, we use the following formula:

$$\text{Sentiment value} \times \text{Number of occurrences}$$

Results

Exploratory Data Analysis and Data Cleaning

In the following sections we investigate the different variables in our dataset that influence our decisions to include certain predictor variables in our final model.

Measure Learning

We create a binary response variable categorizing students' responses to the statement: "this assessment adequately measures my learning" on a Likert scale. We categorize "strongly disagree," "disagree," and "neutral" as a disagreement with the statment. In this research, we are most curious about what constitutes a true agreement rather than a true disagreement. We aim to use this

research to help improve assessment in STEM college courses; therefore, it is critical to understand the difference between true positive sentiment and not.

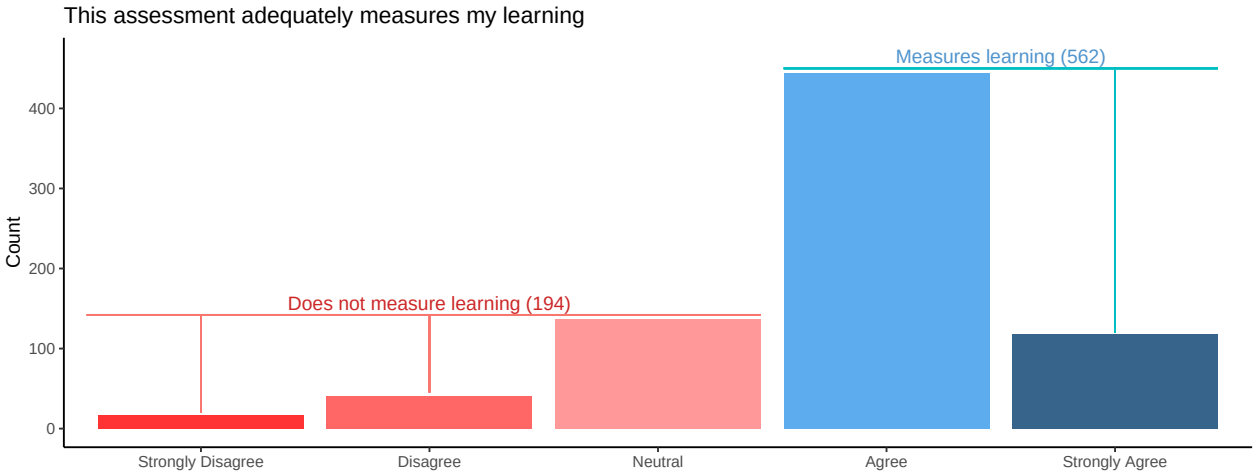

Figure 2: Counts of likert and binary variables for measuring learning

We see in the first plot above (Figure 2), that the majority of students responded "agree" to the survey question. When we group the Likert responses, there are almost three times as many students who report that the assessment does measure their learning compared to students who report that the assessment does not measure their learning. We use the binarized variable of Likert responses in our final model.

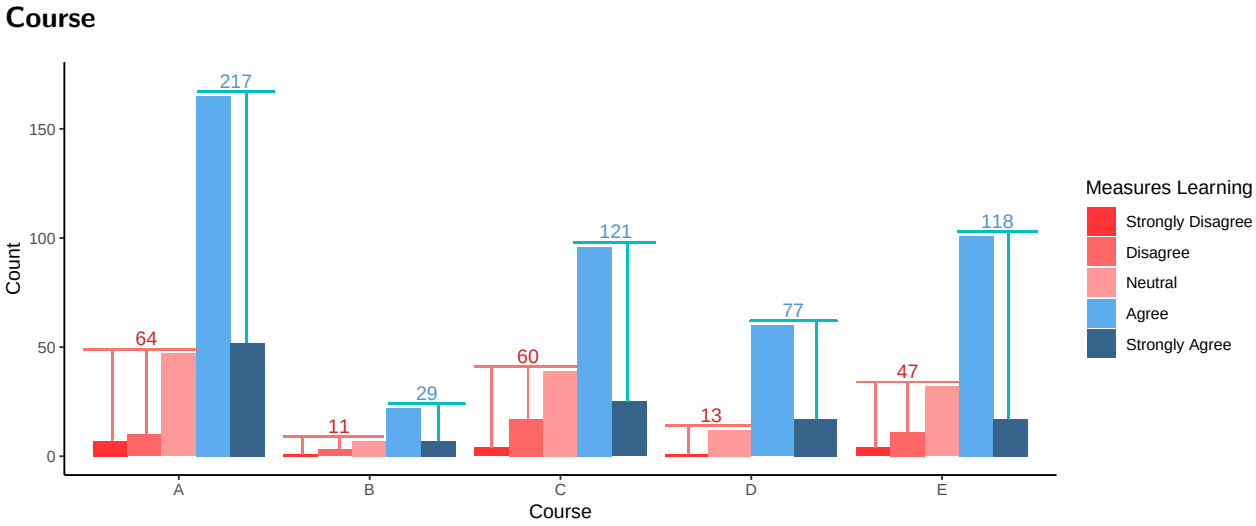

Figure 3: Counts of responses by course

In Figure 3, we see that course B has a smaller number of students than the other courses. For this course, we had to remove survey responses for the second assessment from the data because the grades were given as pass or fail. Therefore, we only have responses for the first assessment for course B, and a smaller sample size as a result. We also see that for class D, no students report that they strongly disagree with the statement that the assessment adequately measures

their learning. In general, the distribution of responses is consistent across courses. In all of the courses, the distribution of responses is centered around "agree", and the majority of students responded "agree" relative to the other Likert options. We use the course variable as a level in our final model.

Assessment Number and Type

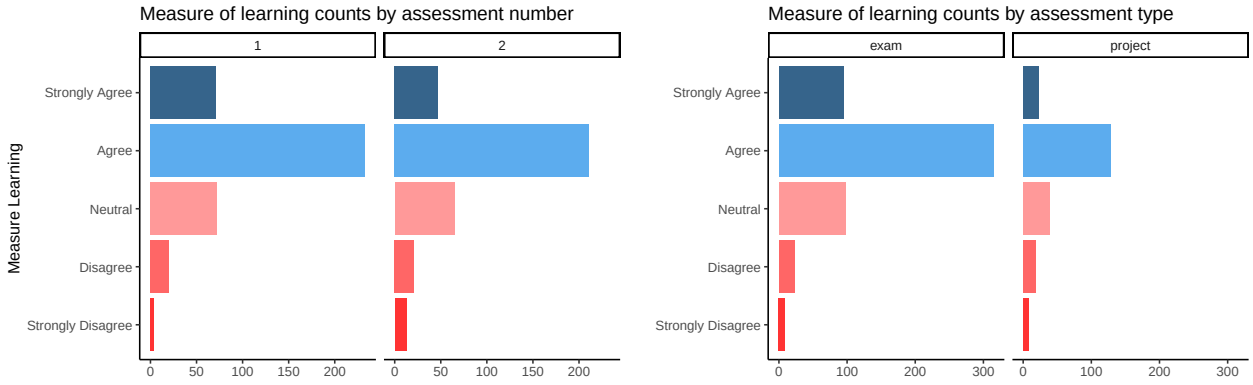

Figure 4: Counts of measuring learning by assesment number and type

In our dataset, assessment number represents the time at which the assessment was taken, first or second. For each course, the first assessment was an exam, and for most courses, the second assessment was a project. In Figure 4, we see that the distributions for likert scale responses by assessment number and assessment type are roughly the same. Since these two variables are highly correlated, we decided to include only one as a predictor in our model. We chose to include assessment type rather than assessment number, because previous research gave us reason to believe that type of assessment may have a greater influence on students' perception of assessment than timing of assessment.

Predicted and Actual Grades

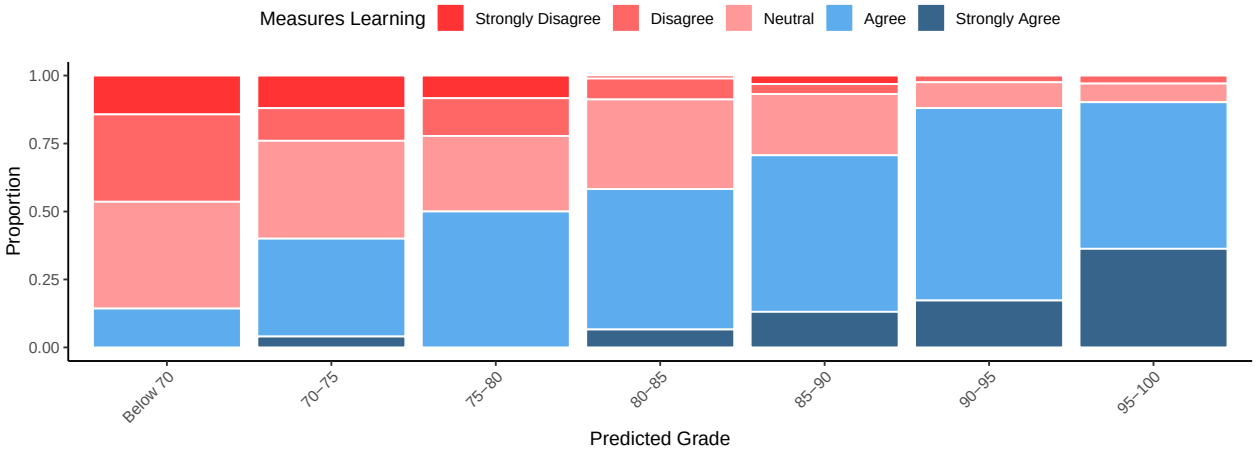

Figure 5: Predicted grades by measure of learning

In Figure 5, we see a strong relationship between students' predicted grades and their perceptions of how well assessments measure their learning. We see that for lower grade predictions, a higher number and proportion of students believe that the assessment does not measure their learning, whereas for higher grade predictions a higher number and proportion of students believe that the assessment does measure their learning. It appears that for every increase in predicted grade range, the proportion of students that agree and/or strongly agree that the assessment measures their learning increases. This gives us reason to believe that the predicted grade may be an important predictor in our model and is included.

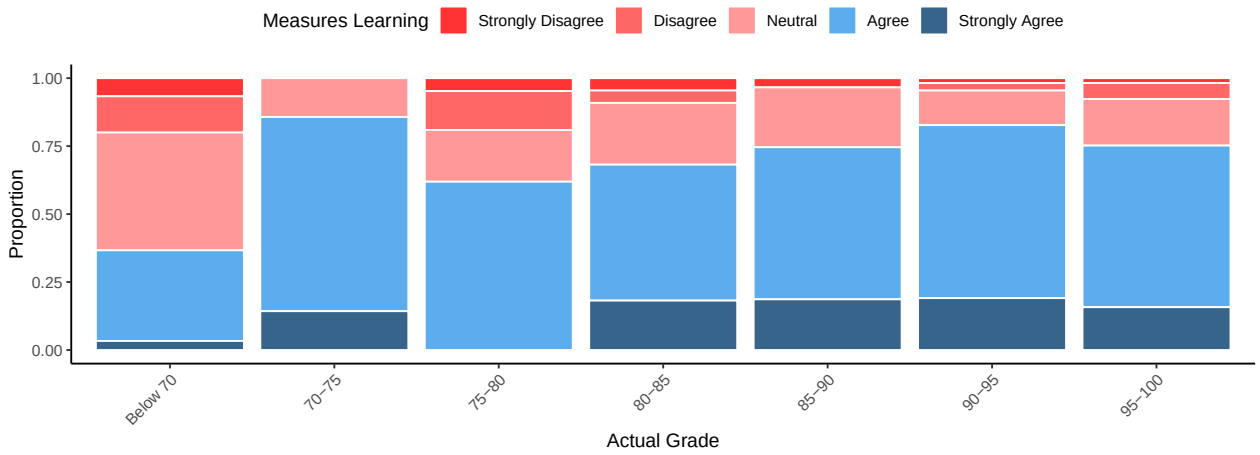

Figure 6: Actual grades by measure of learning

In Figure 6, we do not see a strong relationship between students' actual grades and their perceptions of how well assessments measure their learning. It is interesting to note that most students received grades in the 90-95 and 95-100 grade range. We also see that for every increase in the actual grade range, the proportion of students that agree and/or strongly agree that the assessment measures their learning does not necessarily increase.

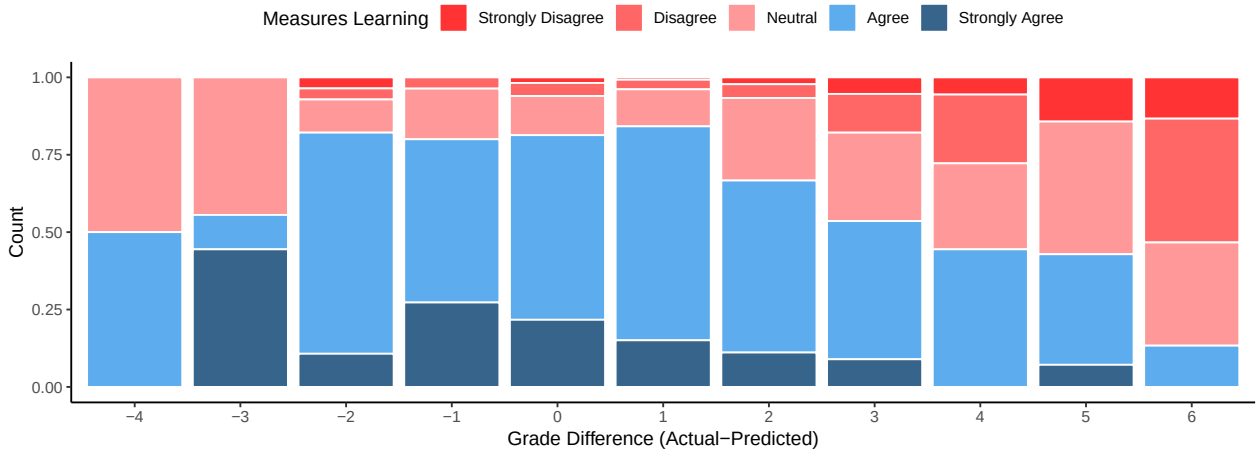

Figure 7: Difference in grades (actual minus predicted) by measure of learning

Based on the analysis above, we decided to explore the difference in actual grades and predicted grades. In Figure 7, we see that the proportion of students who believe that an assessment measures

their learning is higher for smaller grade differences. We also see that students who predicted that they would do poorly on the exam but actually did well on the exam, resulting in a higher grade difference, had a high proportion of students who believed that the assessment did not measure their learning. This EDA suggests that students who inaccurately predict there grades may be more likely to believe that an assessment does not measure their learning.

In our model we consider including actual and predicted grades versus the difference in these grades. In our exploratory data analysis we see a strong relationship between predicted grade and our response variable (Figure 5). Therefore, we have reason to believe that predicted grade may be an important covariate in the model. Although we see a relationship between grade difference and our response variable (Figure 7), predicted and actual grades are more interpretable in the context of our research questions, and therefore we decided to include these individual predictors rather than their difference.

Major

On the first survey, we asked students their first and second major in one question. We used this response for the major variable for the first assessment. On the second survey we asked students their primary major in one question, and their secondary major (if any) in a second question. We used students responses to the first question about their primary major for the major variable for the second assessment.

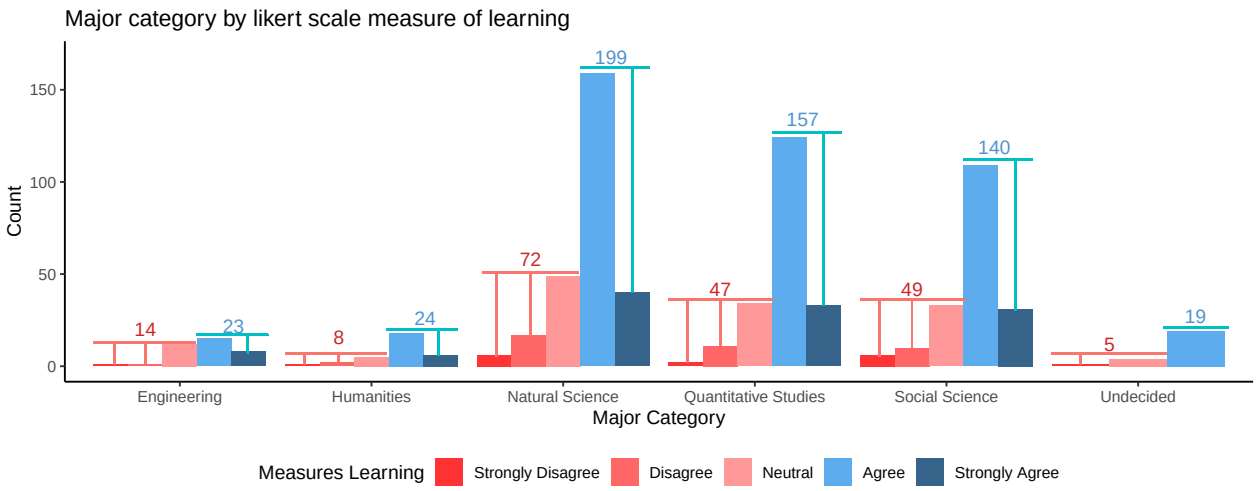

Figure 8: Counts of responses by major category

To clean the major data, we create a major category variable based on the "Areas of Knowledge" (AOK) defined by SCHOOL University, which can be found on SCHOOL's website. Most majors have a one to one mapping with an AOK, however for majors that do not, we used the most common AOK as indicated by the courses for that major on SCHOOL Hub, where all courses and their descriptions are posted. Some students wrote that they were "undecided," which is appropriate given that SCHOOL students do not declare their major until their second year. Furthermore, some students put "Program II," a program which allows students to create their own major. For students who elaborated on a topic were filed into an appropriate AOK category. If there was no additional description, those students were assigned the major category of "undecided." Only two students fell into this category. Furthermore, we collapsed students majoring in subjects in

"Civilizations" or "Arts, Literature, Performance" AOKs into one category, "Humanities" due to the small sample size. The counts within each category can be seen above Figure 8.

In addition to creating a major category variable, we create two binary variables representing major: quantitative studies major or not, STEM major or not. We hypothesize that students majoring in quantitative studies or STEM may be more likely to believe the assessments measure their learning because they may enjoy the subject matter and feel more confident in it. First, we create a variable encoding whether or not the students' major falls in the "Quantitative Studies" AOK. All of the courses surveyed in this study are for majors that fall into Quantitative Studies. Therefore, it would be concerning if students (intending to) major in quantitative studies think that assessments in these courses measure their learning less adequately than students of other majors. If quantitative studies majors believe the assessments do not measure their learning, they may be more likely to turn away from the major despite their initial interest. Second, we create a variable encoding whether or not the students' major is in Science, Technology, Engineering, or Math (STEM). This variable includes majors in Quantitative Studies, Natural Science, and Engineering AOKs. We create this variable for similar reasoning as stated above for the quantitative studies variable. In Figure 9, we see that the counts for these two binary variables are similar. Since these variables are similar, we choose to use the binary variable for quantitative studies in our final model because this variable represents students that (intend to) major in AOKs that overlap with the courses in the study.

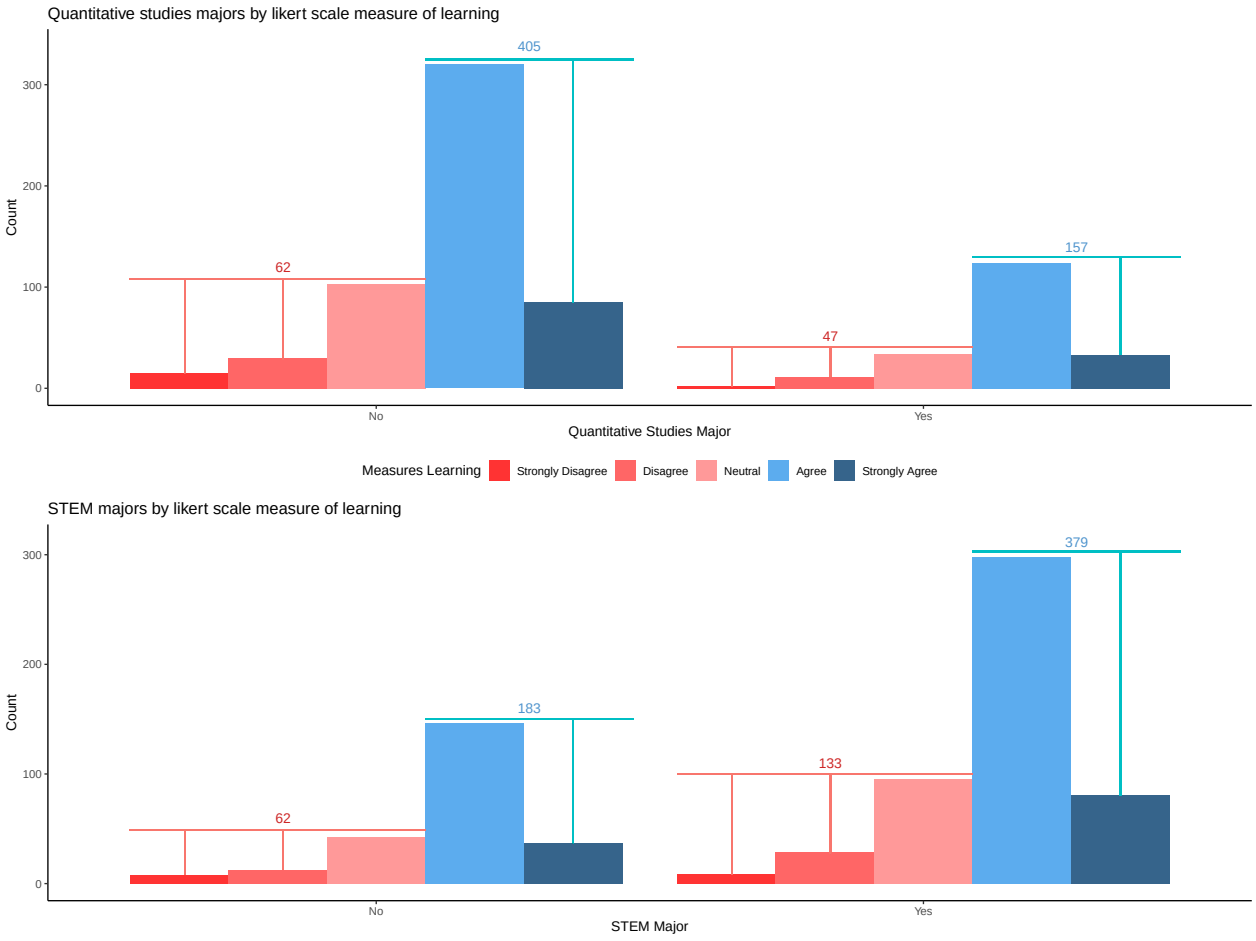

Figure 9: Counts of responses by major

We consider quantifying major in three different formats, in categories of areas of knowledge (AOK) defined by SCHOOL University, binarized as STEM majors or not, and binarized as quantitative studies majors or not. We ultimately decide to include the binary variable for quantitative studies majors in our final model. The major category variable has some categories with relatively small sample sizes, as well as little difference in responses among many of the major categories. Furthermore, we decide to include the quantitative studies variable rather than the STEM variable because all of the courses surveyed are quantitative studies courses. Therefore this variable is more relevant in the context of our research question.

Year

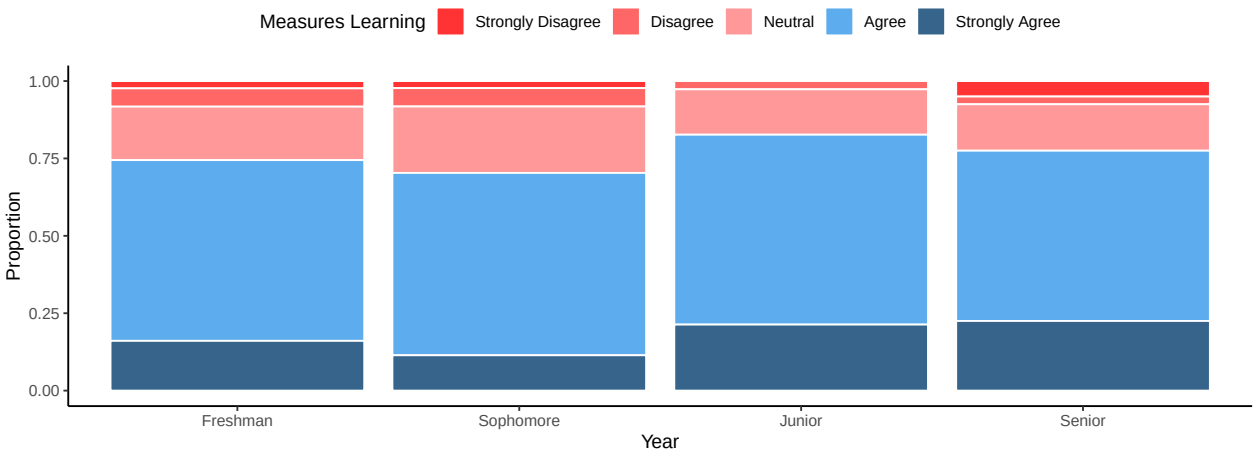

Figure 10: Counts of responses by year

We see in Figure 10, that the majority of students in the data are Freshman. The counts of students decrease as the year in school increases. We expect to see this data given that all of the students surveyed were taking introductory courses. We also see that the proportion of students who agree and strongly agree that the assessment measures their learning is slightly lower for freshman and sophomores. Students in younger grades are more likely to major in STEM or pursue further STEM courses; so we would hope that they tend to believe that assessments adequately measure their learning. Therefore, we include the year variable in our final model.

Model Results

We select variables to include at level one based on our exploratory data analysis and what aligns with our research question. In our final model, the response variable is whether students believe an assessment adequately measures their learning or not, and the predictors include assessment type, predicted grade, actual grade, whether the student is a quantitative studies major or not, and year in school. The levels of the model include the course and student. The model results can be found below.

With regard to random effects, the model results report that the variance in intercepts from class-to-class is approximately zero. This suggests that the odds that a student believes an assessment

measures their learning do not vary by class. The fixed effects from the model results can be seen in Table 5 below.

Table 5: Multilevel Logistic Regression Fixed Effects Estimates

Variable	Estimate	SE	95% CI	p-Value
Intercept	-1.92	0.65	(-3.2, -0.65)	0.00
Project	-0.52	0.23	(-0.98, -0.07)	0.02
Predicted 70-75 Grade	1.37	0.70	(-0.01, 2.74)	0.05
Predicted 75-80 Grade	2.03	0.67	(0.72, 3.34)	0.00
Predicted 80-85 Grade	2.16	0.62	(0.95, 3.37)	0.00
Predicted 85-90 Grade	2.84	0.61	(1.64, 4.04)	0.00
Predicted 90-95 Grade	4.00	0.62	(2.79, 5.22)	0.00
Predicted 95-100 Grade	4.19	0.68	(2.85, 5.52)	0.00
Actual 70-75 Grade	2.16	1.24	(-0.28, 4.59)	0.08
Actual 75-80 Grade	0.34	0.64	(-0.92, 1.59)	0.60
Actual 80-85 Grade	0.30	0.66	(-0.99, 1.6)	0.64
Actual 85-90 Grade	0.07	0.55	(-1.02, 1.15)	0.91
Actual 90-95 Grade	0.19	0.53	(-0.85, 1.23)	0.72
Actual 95-100 Grade	0.26	0.48	(-0.67, 1.19)	0.59
Quantitative Studies Major	0.14	0.22	(-0.29, 0.57)	0.53
Sophomore	-0.45	0.21	(-0.87, -0.03)	0.04
Junior	0.04	0.35	(-0.66, 0.73)	0.92
Senior	-0.14	0.45	(-1.02, 0.75)	0.76

We see from Table 5 that when all predictors are at the baseline, meaning that the type of the assessment is an exam, the predicted grade is below 70, the actual grade is below 70, the student is a freshman, and the student is not majoring in quantitative studies, the odds that a student believes the assessment measures their learning are 0.15, i.e. a there is a 14% likelihood that a baseline student believes an assessment measures their learning.

Moreover, we see interesting model results regarding grade prediction. In Figure 11, we see a general trend that when students predict higher grades, the log odds that a student believes an assessment measures their learning increase. For example, when a student predicts their grade to be in the 70-75 range, the odds that a student believes the assessment measures their learning are 3.93 times the probability of when a student predicts their grade to be below 70, holding all else constant, versus when a student predicts their grade to be in the 95-100 range, the odds that a student believes the assessment measures their learning are 65.72 times the probability of when a student predicts their grade to be below 70, holding all else constant.

The estimates for actual grade predictors, however, do not follow the same trend. When a student's actual grade is in the 70-75 range, the odds that a student believes the assessment measures their learning are 8.63 times the probability of when a student's grade is below 70, holding all else constant, versus when a student's actual grade is in the 95-100 range, the odds that a student believes the assessment measures their learning are 1.3 times the probability of when a student's actual grade is below 70, holding all else constant. Although the estimates for the actual grade predictors are all positive, the log odds that an assessment measures a student's learning do not increase as the actual grade increases.

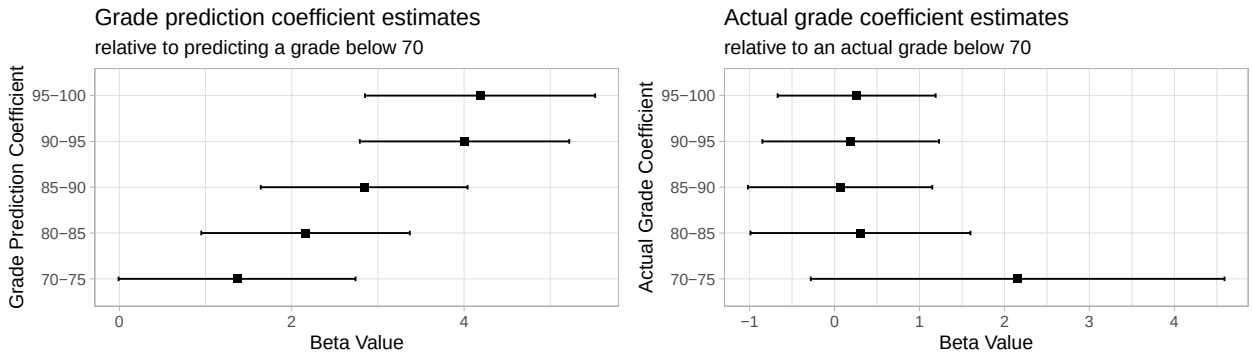

Figure 11: Predicted and actual grade model estimates

When a student (intends to) major in quantitative studies, the odds that the student believes the assessment measure their learning are 1.15 times the probability of when a student does not (intend to) major in quantitative studies, holding all else constant. In other words, a student who (intends to) major in quantitative studies is 15% more likely to believe that an assessment reflects their learning that a student who (intends to) major in non-quantitative studies. However, we see that the 95% confidence interval for this estimate contains zero, therefore it is possible that this estimate is actually negative and not positive.

Moreover, when a student is a sophomore, the odds that the student believes the assessment measures their learning are 0.64 times the probability of when a student is a freshman, holding all else constant. In other words, sophomores are 37% less likely to believe that an assessment reflects their learning than a freshman.

In Figure 12, the ROC curve for the model reports an AUC of 0.798, meaning that there is a 79.8% likelihod that the model ranks a random success (the student believes an assessment measures their learning) more highly than a random failure (the student believes an assessment does not measure their learning). This AUC score suggests that the model does a reasonable job classifying whether students believe an assessment reflects their learning.

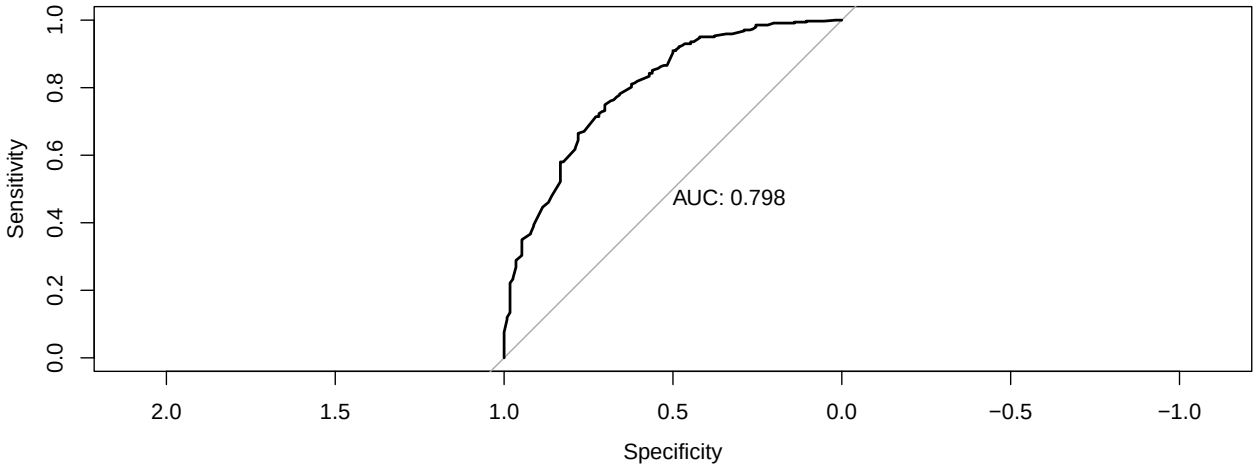

Figure 12: ROC Curve

Text Analysis Results

Of the 757 survey responses included in our model results, students provided text responses in 689 or 91% of the surveys. Students were asked to explain their Likert response to the statement: "this assessment adequately measures my learning." There were roughly 24 words per text response on average. We evaluate sentiment using sentiment scores, which are defined by the Bing lexicon in a binary fashion as positive or negative. An example of a response with a more negative sentiment score was, "I was super stressed while taking the exam, and felt that anxiety and stress affected my performance a lot." Conversely, an example of response with more positive sentiment score was, "effectively tested all skills learned to an appropriate level of difficulty."

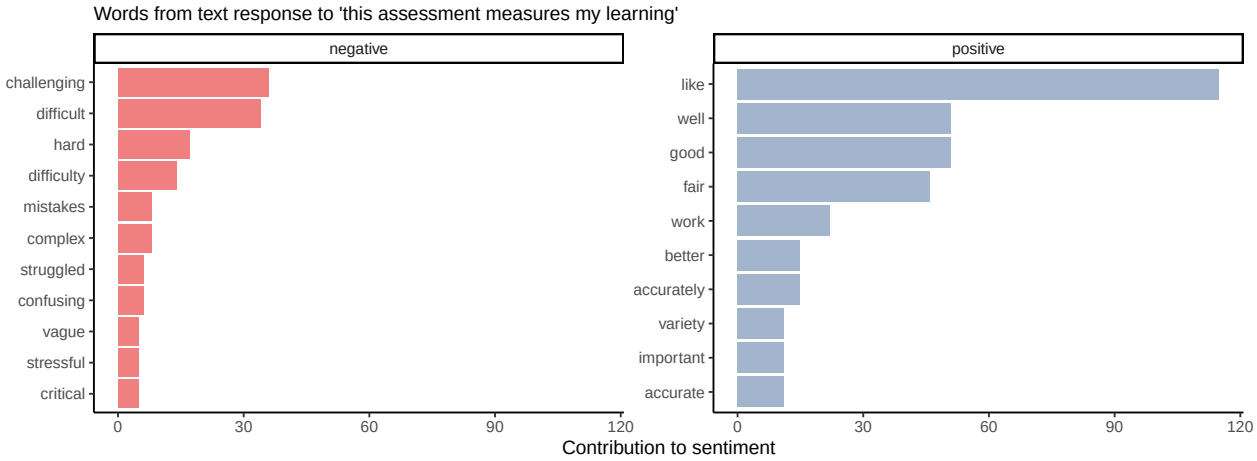

Figure 13: Positive and Negative Bing Sentiment Word Counts

In Figure 13, we see the most common positive and negative words that appear in the text responses. We use sentiment scores from the Bing lexicon (Silge and Robinson (2016)). We see that some of the most common negative words include "challenging" and "difficult" which have clear negative sentiment. Some of the words that initially appeared on this list such as "problem" and "problems" may have different sentiment given the context of the text responses. Therefore, after further investigation, we see that "problem" and "problems" are most commonly used in a neutral context to describe things like "real-world problems" and "problem solving" rather than having a problem. We see similar issues with the most common positive words as well. Words like "well" or "good" or "fair" have clear positive positive sentiment; however, we hypothesize words like "pretty" and "enough" that appeared on this list may have more neutral sentiments in the context of the text. In the text, most students used "pretty" as an adverb to describe both positive and negative words such as "pretty fair" and "pretty challenging." Similarly, "enough" is used in both positive and negative contexts such as "fair enough" and "not enough." Previous papers about sentiment analysis such as Atteveldt, Velden, and Boukes (2021) suggest that models do a nice job assessing sentiment, however, the best overall classification performance is attained with "trained human or crowd coding" and therefore, we have justification for classifying sentiment by hand. Thus, with these commonly appearing words, we classify them with neutral rather than positive or negative sentiment. It is important to note, however, that we did not check the classification for every word in the text, and thus it is possible that the Bing sentiment scores do not accurately classify all of the words in the data. However, we do not have a strong reason to believe that the sentiment is inaccurately represented for the majority of the text.

The miscalculation of the sentiment of some of the words in the text, lead us to investigate bigrams in the responses. Most bigrams that do not contain "stop words" appear less than twenty times in the text. Thus, it is difficult to infer any major trends in bigrams.

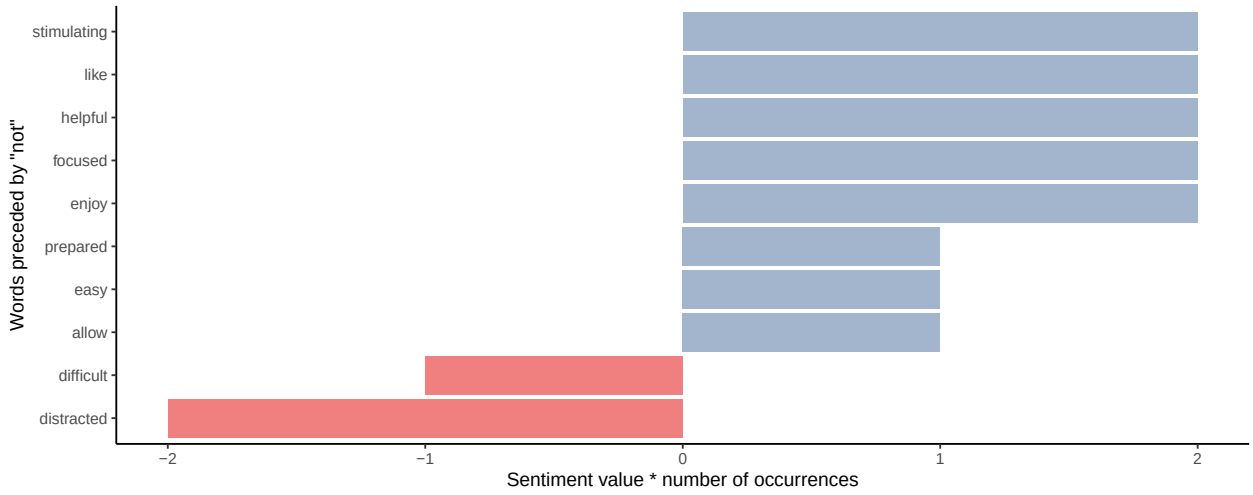

Figure 14: Bigrams starting with 'not'

For example, we look at bigrams including negation words. We use the AFINN lexicon to calculate sentiment scores which are summed across both words in the bigram. The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. In Figure 14, we see words preceded by 'not' that have the greatest contribution to sentiment values, in either a positive or negative direction. There are multiple bigrams like 'not stimulating' and 'not like' which may lead to misidentification and may make students explanations for their survey responses seem much more positive than it is. We can also see phrases like 'not difficult' and 'not distracted' sometimes suggest text is more negative than it is. However, none of these bigrams occur frequently in the text, so we do not have reason to believe that the sentiment analysis results are substantially misrepresented.

Measure Learning

In Figure 15, we see that the sentiment scores increase as the Likert scores increase in agreeability with the statement 'this assessment adequately measures my learning.' We only see a negative sentiment score among students who strongly disagree with the statement. The rest of the Likert scores result in text responses with positive sentiment. These results are in line with what we would expect for Likert responses.

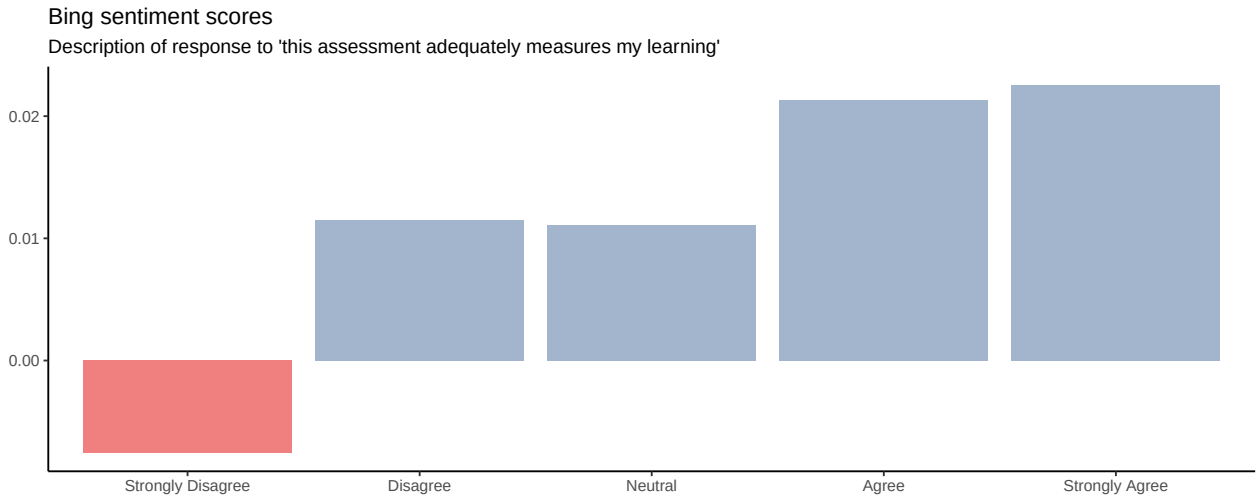

Figure 15: Response Variable Sentiment Analysis

Assessment Type

In Table 6, we see that the word counts for exams are higher than for projects which makes sense given the smaller number of students who were surveyed about a project. We see that the ratio of negative words in the text for both exams and projects is approximately the same, with approximately 1.5% of the words in the text responses for exams and projects being negative. We see that the sentiment score for exams is slightly larger than for projects, at 0.020 versus 0.014, which we would expect given the model results.

Table 6: Bing word sentiment counts by assessment type

Type	Positive	Negative	Neutral	Total	Negative Ratio	Sentiment Score
exam	381	163	10430	10974	0.015	0.020
project	162	86	5335	5583	0.015	0.014

Predicted Grade

In Figure 16, on the left we see that ratio of negative words in the text responses tends to decrease as the predicted grade category increases, with the exception of the 75-80 predicted grade category. Similarly, on the right, we see that the sentiment score of the text tends to increase as the predicted grade category increases. We would expect these sentiment analysis results given the results from our modeling which suggest that the odds of believing an assessment measures learning increase as the predicted grade category increases. Therefore, these text analysis results further support our model results.

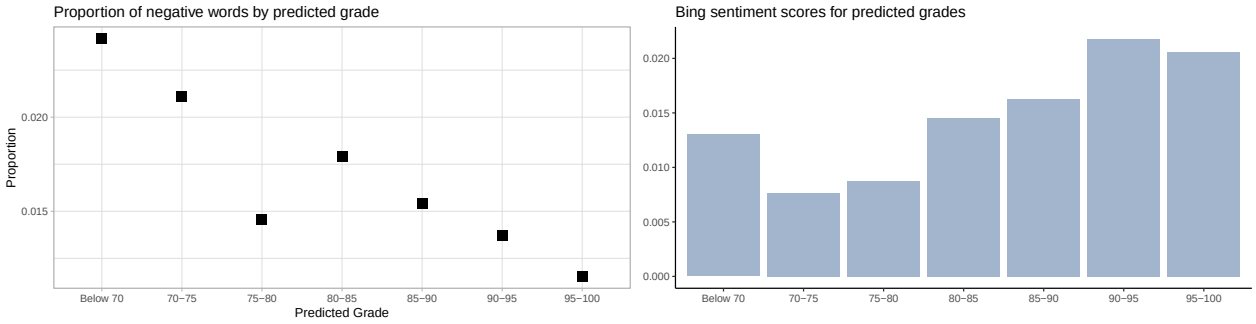

Figure 16: Predicted Grades Sentiment Analysis

Actual Grade

In Figure 17, we see a similar trend with the sentiment of actual grades to predicted grades. On the left, we see that the ratio of negative words tends to decrease as the actual grade range increases. Similarly, on the right, we see that the sentiment score tends to increase as the actual grade increases.

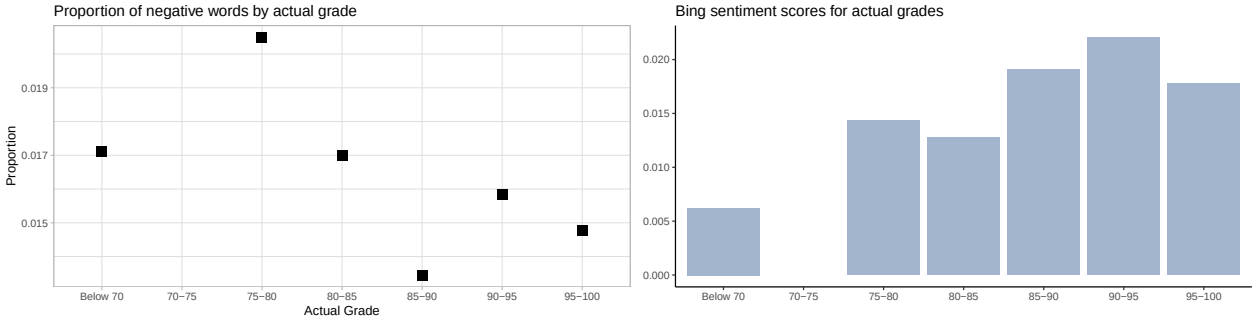

Figure 17: Actual Grades Sentiment Analysis

Discussion & Conclusion

In this paper, we aim to quantify relationships between student perception of assessment and student identity, class format, and grading in introductory quantitative studies courses at SCHOOL University. If we better understand students' perception of assessment, we can better curate assessments, a key measure of academic achievement, that motivate students to learn actively and pursue STEM.

From our research we highlight three main takeaways. First, we see that students are more likely to perceive that exams measure their learning than projects. Before conducting this research, we hypothesized the opposite, that students would be more likely to perceive that projects measure their learning than exams because previous research suggests that project based learning has a positive impact on students' motivation to attend and participate in their courses (Klegeris and Hurren 2011). Therefore, we found this result surprising, especially given the recent push to diversify assessment formats at the university level. Since students tend to believe that exams better measure their learning, instructors may want to reconsider equating projects with exams in evaluating academic achievement.

Second, we see that students are more likely to perceive that assessments measure their learning when they predict they will receive a higher grade. We see a strong relationship between predicted grades and students perception of assessment with high confidence. This suggests that students perception of assessment may be dependent on the grade they think they will receive. This highlights that student learning may be driven by grades more than learning. Going forward, students may learn better or at least understand their learning if assessments are not so tied to grades. This research provides evidence to suggest that we should put less emphasis on grading and more emphasis on learning in the education system.

Third, we see that quantitative studies majors are more likely to believe quantitative studies assessments measure their learning than students of other majors. However, given the lack of statistical significance of this estimate, we cannot claim that the true estimate is positive with high confidence. Furthermore, an important caveat to note is that the positive estimate may be a result of the fact that quantitative studies majors may feel more confident in these courses in general because they are quantitative studies courses. Nonetheless, although we ideally want all students to feel that assessments measure their learning, it is especially important that students who major or intend to major in quantitative studies courses believe that introductory quantitative studies courses' assessment adequately measure there learning because they will be more motivated to continue down this STEM track. Going forward, instructors should continue to monitor the churn rates in the majors of their corresponding courses. It is important for students who believe they are interested in quantitative studies remain interested given the grave demand for STEM workers.

A potential additional takeaway for researchers interested in conducting similar studies is that it may not be necessary to collect text data. Firstly, it can be difficult to collect quality text data, and participation rates may be higher for shorter surveys without text questions. Moreover, in our research, we see that our text analysis supports our other findings, but does not provide enlightening information that answers our research question. The sentiment of text responses were as expected, matching the sentiment of Likert responses (i.e. students who strongly disagree were negative in their sentiment, and students who strongly agree were positive in their sentiment). Therefore, we believe that this study would be adequate without any text analysis, and similar studies should highly consider the trade offs in asking text questions on surveys.

Ultimately, our research highlights multiple interesting findings, however, there are still multiple limitations to this study. First, in our data collection, one of the courses did not administer a project as a major assessment in the course, so students were surveyed on two exams. In another course, students were surveyed on an exam and a project, however the project was graded with pass or fail rather than points. Thus, the second set of survey responses for this course was thrown out. As a result, there were only 217 survey responses about a project, whereas there were 540 survey responses about an exam. Furthermore, we do not account for assessment differences beyond type (exam versus project). However, the formatting of the exams and projects varies by course. For example, some exams were administered during an in-person class period, whereas other exams were administered online and open for multiple days. With the projects, some of them were individual projects and some of them were small group projects. In our research, we did not consider other predictors regarding the assessment type in our model which might have accounted for some variability in the results.

Evidently, there is still much to be explored regarding students' perception of assessment in STEM. In our research, we see that the type of assessment as well as students' predicted grade impact how well they believe an assessment measures their learning. Therefore, in the future, it would be interesting to control for traditional (in person, hand written, timed) exams versus other forms

of alternative assessment and evaluate students' perception. With an even better understanding of which assessments measure students' learning best, instructors can improve their courses to boost student motivation. It would also be interesting to control for courses with and without (or pass/fail) grades or to compare how different styles of grading impact students' perception of assessment. In our research, it seems that are influenced and motivated by grades, so it would be helpful to understand students' perceptions when grades play a lesser role in their learning. It would also be interesting to expand the study to include a wider range of STEM courses from different universities in an effort to account for a more representative study group. Moreover, it would be informative to follow students over time and track their perception of assessment from introductory to advanced course and see which students pursue a career in STEM after graduation. Finally, in our research we do not collect any demographic data. The STEM gap and inequalities in STEM are a widely discussed topic, so it would be interesting to run the same study and investigate student perception of assessment conditioned on different student demographics that we care about.

Appendix

Appendix A: Informed Consent

Page 1

- Are you 18 years or older?
  - Yes (If yes, participant is taken to Page 2).
  - No (If no, survey ends for participant and they see the following: "Thank you for your interest this study. We can only ask consent from those who are 18 years or older. Please come back after you turn 18 to complete this consent form.").

Page 2

We are learning too!

Key Information

You are being asked to participate in a research study being conducted by undergraduate student in Statistical Science & Computer Science, NAME, supervised by NAME, Assistant Professor of the Practice at SCHOOL University. The purpose of this study is to: understand both students' and professors' perceptions of assessment and whether they are aligned and/or effective. The primary question we are trying to answer is: how do students measure different forms of assessment as reflections of how well they learned in introductory quantitative courses? We are asking you to share your opinions via two surveys and share your grades on these two assessments in binned ranges.

Voluntariness and Confidentiality

Your participation is completely voluntary, and you may withdraw at any time. You do not have to agree release the surveys used in this research. Your instructor will not know who consents to participate in the research. Your decision will have no impact on your grades in this or any other course you have taken or will take in the future.

Identifying information will be used to remove data from students who elect not to participate. Data will not be made public or used for future research purposes.

Survey and Release of Course Data for Research

We are asking that you complete two short surveys (approximately five minutes each) throughout the semester and share your grades on two assessments in binned ranges to be used in research. Your data will be used to help us better understand the teaching and learning experiences here at SCHOOL University. We are absolutely not making any judgments about any individual participants. All data will be deidentified by SCHOOL Learning Innovation before made accessible to the research team. None of your personal information will be included in any analysis or publications of our findings. Please select one of the following:

- Yes, I agree to release my survey responses and grades to be used for research purposes
- No, my survey responses and grades CANNOT be used for research purposes

Contact Information

For questions about this research, please contact NAME at EMAIL and NAME at EMAIL.

If you agree to be in this study but later change your mind and want to withdraw, you can return to this survey any time during the semester and change your response. You may also contact SCHOOL Learning Innovation at learninginnovation@SCHOOL.edu and ask to be removed from the research.

For questions about your rights as a participant in this research, please contact the SCHOOL Campus IRB at campusirb@SCHOOL.edu with reference to Protocol #[2022-0545]. Please print this page if you would like a copy for your records.

Appendix B: Model Conditions

First, because we use a binomial distribution, we check if there is no over or underdispersion in your model.

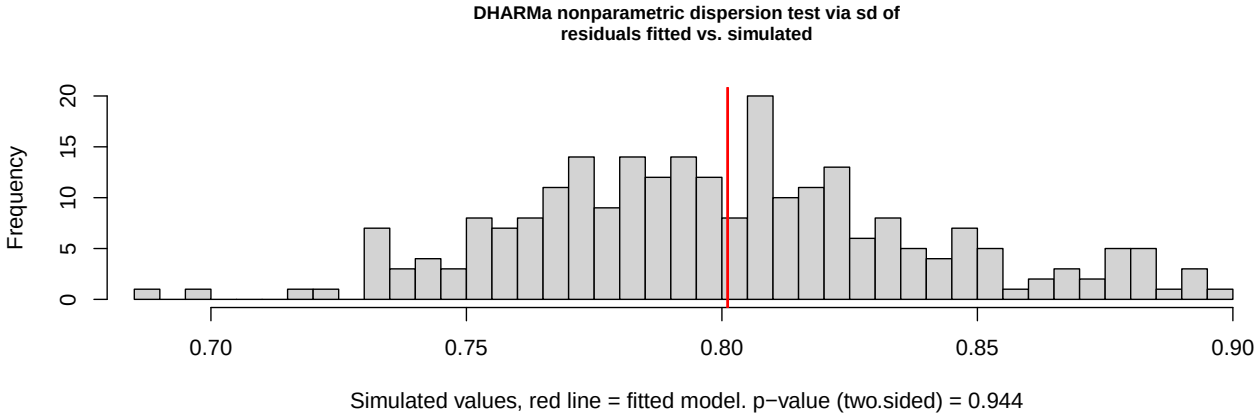

Figure 18: Model Conditions Tests

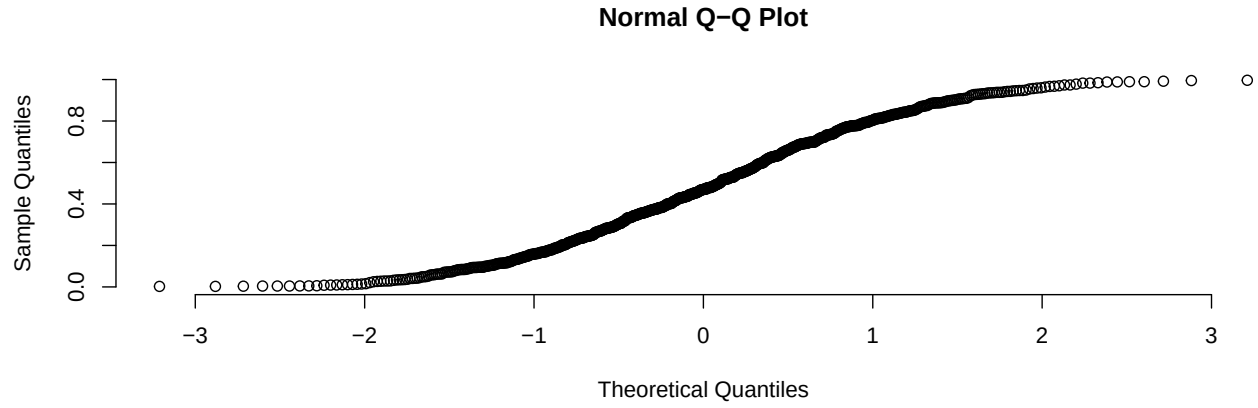

Figure 19: Model Conditions Tests

We use several overdispersion tests from the `DHARMa` package that compare the dispersion of simulated residuals to the observed residuals. The nonparametric disperison test via standard deviation of residuals versus simulated does not show any reason for concern with the model. Furthermore, there is no detectable pattern in QQ plot. Therefore, we have no reason to believe that there is overdirspersion or underdispersion in the residuals and we assume the conditions are met.

Table 7: Multicollinearity: VIF for Model Predictors

	GVIF	Df	GVIF^(1/(2*Df))
assessType	1.33	1	1.15
gradePred	1.51	6	1.03
gradeActual	1.87	6	1.05
qs	1.06	1	1.03
year	1.21	3	1.03

Since each of the VIF values for the predictor variables in the model are close to 1, multicollinearity is not a problem in the model.

References

Atteveldt, Wouter van, Mariken A. C. G. van der Velden, and Mark Boukes. 2021. "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms." *Communication Methods and Measures* 15 (2): 121–40. https://doi.org/10.1080/19312458.2020.1869198.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Bidwell, Allie. 2015. "More Students Earning STEM Degrees, Report Shows." *U.S. New and World Report*. https://www.usnews.com/news/articles/2015/01/27/more-students-earning-degrees-in-stem-fields-report-shows.

Brown, Patrick L., James P. Concannon, Donna Marx, Chris Donaldson, and Alicia Black. 2016. "An Examination of Middle School Students' Stem Self-Efficacy, Interests and Perceptions." *Journal of STEM Education: Innovations and Research* 17 (3). https://www.jstem.org/jstem/index.php/JSTEM/article/view/2137.

Klegeris, Andis, and Heather Hurren. 2011. "Impact of Problem-Based Learning in a Large Classroom Setting: Student Perception and Problem-Solving Skills." *Advances in Physiology Education* 35 (4): 408–15. https://doi.org/10.1152/advan.00046.2011.

Labor Statistics, U. S. Bureau of. 2022. "Employment in STEM Occupations." *U.S. Bureau of Labor Statistics*. U.S. Bureau of Labor Statistics. https://www.bls.gov/emp/tables/stem-employment.htm.

Lynch, Sharon J., Erin Peters-Burton, and Michael Ford. 2014. "BUILDING STEM Opportunities FOR ALL." *Educational Leadership* 72 (4): 54–60.

Richardson, John T. E. 2005. "Instruments for Obtaining Student Feedback: A Review of the Literature." *Assessment and Evaluation in Higher Education* 30 (4): 387–415. https://doi.org/10.1080/02602930500099193.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in r." *JOSS* 1 (3). https://doi.org/10.21105/joss.00037.

White, Erin, and Ariana Shakibnia. 2019. "State of STEM: Defining the Landscape to Determine High-Impact Pathways for the Future Workforce." *Proceedings of the Interdisciplinary STEM Teaching and Learning Conference* 3 (1). https://doi.org/10.20429/stem.2019.030104.

Will, Madeline. 2022. "They Recruited 100,000 Stem Teachers. Now They're Setting Their Sights Even Higher." *Education Week*. Education Week. https://www.edweek.org/leadership/they-recruited-100-000-stem-teachers-now-theyre-setting-their-sights-even-higher/2022/09.