Available online at ijci.wcci-international.org

International Journal of Curriculum and Instruction 17(1) (2025) 101–132

IJCI
International Journal of Curriculum and Instruction

Students' Perceptions of Learning Outcomes: Traditional Versus Growth-Based Grading in Biology Education

Hannah Kinmonth-Schultz [a] *, Kinsey Simone [b]

[a,b] *Tennessee Technological University, 1 William L Jones Dr, Cookeville, TN 38505, USA*

Abstract

This study examined whether differences existed in college students' perceptions of improvement in critical thinking, content knowledge, and STEM motivation after taking a biology course and receiving traditional-based or growth-based grading methods on scientific write ups. The quasi-experimental study used data collected from a reflection survey during in 2023 from a public Tennessee university. Analyses included factor analysis, a MANCOVA, and sentiment analysis. The interaction of group and gender was significant, with men perceiving overall higher improvements in content knowledge, and women demonstrating greater STEM motivation than men in the treatment group. Students had overall positive sentiments, with those in the treatment group emphasizing growth. This study highlights a positive association between growth-mindset interventions and student outcomes. Implications are provided.

Keywords: traditional grading, growth-based grading, critical thinking skills, Biology, scientific writing

1. Introduction

The purpose of the current study was to assess differences in perceptions of college students in a biology course based on whether they received growth-based grading or traditional grading and controlling for gender. Assessed perceptions included improvement in critical thinking skills, content knowledge, and motivation in STEM. The study used a quasi-experimental static control group post-test design, and instruments included students' responses on a researcher-developed reflection survey. A review of current relevant literature is followed by methods, results, conclusions, and implications for future research. The university's Institutional Review Board granted approval for the study.

* Corresponding author: Hannah Kinmonth-Schultz. ORCID ID.: https://orcid.org/0000-0003-3541-007X
E-mail address: hkinmonth@tntech.edu

1.1. Background

Despite much focus on the gender and diversity gap in STEM fields over the last several years, the percentages of women, members of underrepresented groups, and students who are the first in their families to attend college still lag behind the representation of those same groups in the general population (Hamrick et al., 2021). Phenomenon such as "belonging uncertainty", "imposter syndrome", or "stereotype threat" contribute to this gap in representation, as members in underrepresented groups may perceive, as inherent reflections of their ability, failures experienced in fields for which they do not have a strong identity connection (Casad & Bryant, 2016). This internalization of failure can result in individuals avoiding opportunities for feedback to improve and to otherwise disengage from the work and school environments (reviewed in (Casad & Bryant, 2016)). In the school environment, disengagement could result in students missing class, skipping answers on exams, or failing to turn in assignments (Davies et al., 2005; Koenig et al., 2011; Cheryan et al., 2011; Dee, 2014; Dweck & Yeager, 2019).

Growth mindset interventions, which emphasize the potential for students to improve in their ability, have proven to be effective and were associated with stronger test scores across multiple studies, especially for students with otherwise greater likelihood of poor performance (Yeager & Dweck, 2020). Individuals with a growth mindset show greater resiliency and are less likely to shy away from challenges (Dweck & Yeager, 2019). Yet, while effective programs exist, there is still a need to assess how best to help students internalize the growth mindset, especially if they come to the classroom with preconceived perceptions of 'belonging' in certain groups or career tracks.

Facilitated mentor/mentee relationships between students in underrepresented groups and faculty members or peers have proven to be successful, low-cost interventions having a large impact on student success (Campbell et al., 2012; Dennehy & Dasgupta, 2017). For example, students participating in same-gender peer mentoring interventions for freshmen and sophomore females in STEM helped mentees to gain more self-efficacy, greater success in engineering, and more motivation to pursue STEM fields (Dennehy & Dasgupta, 2017). These mentor/mentee relationships are likely to be successful in large part because mentors help to establish that failures such as low grades on assignments or exams are normal and, thus, can be overcome (Brady et al., 2020)—reinforcing both belonging and a growth mindset. In fact, facilitated mentor/mentee interactions of just 1 hour that focused specifically on normalization of failure had positive impacts on perceptions of life and career satisfaction in participants over 10 years later (Brady et al., 2020). However, students tend to seek out mentors that are the most accessible—whether these be faculty, staff, or peers (Campbell et al., 2012). Additionally, students, especially those struggling with self-confidence, often do not reach out to mentors unless required to do so. Therefore, programs that facilitate and require mentor/mentee interactions that

specifically incorporate a STEM focus, are more likely to be successful at closing the representation gap in STEM. Entry-level, general biology courses offer unique opportunities to institute growth mindset interventions that reach a broad swath of the student body as these courses are taken by students interested in professional, applied, and research career tracks. However, these classes are often large, limiting opportunities for individualized attention, and increasing the likelihood that those experiencing belonging uncertainty will disengage after experiencing early failure or avoid difficult tasks at the outset to limit the negative feedback they receive (Casad & Bryant, 2016). Therefore, interventions that require involvement may be more successful than those that do not.

Reading scientific literature and interpreting data are tasks that are overwhelming for many students, as is concisely and accurately conveying scientific ideas through writing. Thus, incorporating these activities into an introductory biology class provides an opportunity to incorporate and test the effectiveness of strategies designed to promote a growth mindset. In the current study, the researchers instituted a series of three skills-based activities in the semester called "write-ups" in which students interpreted a data figure and relayed the information through writing geared towards a general scientific audience.

1.2. Purpose of Current Study

This study compared the perceptions of improvement in critical thinking skills, content knowledge, and motivation in STEM of students in a Biology course who received either traditional-based grading methods or growth-based grading methods. The following research questions were explored:

1. Are there differences in students' perceptions of course outcomes (critical thinking skills, content knowledge, and motivation in STEM) based on whether they received traditional versus growth-based learning?
2. Are the differences, if any, in students' perceptions of course outcomes (critical thinking skills, content knowledge, and motivation in STEM) based on whether they received traditional versus growth-based learning the same for men and women or non-binary individuals?
3. Are there differences in proportions of positive versus negative qualitative sentiments of students towards a biology course depending on whether they received traditional versus growth-based learning?

2. Methods

This quantitative study used a quasi-experimental static control group posttest design. Factor analysis was used to identify underlying constructs within collected survey data, and a multivariate analysis of covariance (MANCOVA) was conducted to determine differences in three identified constructs, including critical thinking skills, content knowledge, and motivation in STEM, controlling for gender. Analysis of covariance (ANCOVAs) were conducted on each dependent variable using a Bonferroni adjustment to examine significant differences across group and genders after the MANCOVA. Sentiment analysis was then conducted on participants' qualitative responses. Descriptions of the setting and participants, data collection methods, instrumentation, and data analysis follow.

2.1. Setting & Participants

Participants included 81 students enrolled in a General Botany course at Tennessee Technological University (Tech), an intermediate-sized, public university in Southeastern Tennessee that enrolls approximately 10,000 students annually, of which about 85% were undergraduate students in Fall of 2022 ('About Tennessee Tech – Facts and Figures', 2023). Annually, tech graduates 400 to 600 students in STEM majors on average. Tech is a regional university located in Cookeville, TN that serves the rural Upper Cumberland region. 93% of students were In-State students according to the Fall 2022 census. Tech serves an economically distressed rural area, in which five of the surrounding counties are on Tennessee Governor Bill Lee's list of 10 economically distressed counties in the state, making them among the 10 percent most economically distressed counties in the nation. An additional 10 are on Lee's list of 32 at-risk counties ('Transparent Tennessee', 2023).

This course has been a required course for students in the Biology and Wildlife Fisheries Science majors, including those seeking professional degrees in medically-related fields. It is also part of the general education curriculum, as a biology elective, attracting students in the Engineering and Computer Science Majors. As such, the course includes students interested in a range of STEM fields. More than 70% of students tend to be second-year students; however, the course does attract some students in first, third, and fourth year or greater as well. Of the 78 students included in the analysis, 41 identified as women/nonbinary persons (49.4%), and 37 identified as men (44.6%). Originally, 1 participant identified as nonbinary and was included with the gender category of women to be inclusive of all participants. Three participants preferred not to disclose their gender and were not included in the analysis. Based on their lab section and out of eight sections, students were randomly assigned to treatment or control and to small groups used in lecture. This was to ensure that the students most likely to talk about grades and assignments received the same treatment.

2.1. Measurement & Instrumentation

After random assignment, students were given a syllabus corresponding to the control or treatment group as applicable, which explained their grading scheme (either growth- or traditional-based). Regardless of treatment, all students received three write up assignments throughout the semester, in which they responded to a data figure. Write ups were composed of five, one-to-four sentence sections that included the following: background, question, methods, observations, and conclusions. Write-ups were graded on a rubric that assessed each section as well as overall writing quality (See Appendix A). As this was a skills-based assessment, the researchers expected students to improve over the semester providing a framework to incorporate growth-mindset language. Specifically, the course instructor told students within the treatment group that, "scientific writing is hard, interpreting scientific data is hard, and that we expect students to struggle in the beginning, but that with practice, they will get better." Additionally, because growth was expected, the instructor told students within the treatment group that the rubric scores they received on the later write-ups would replace earlier rubric scores. To encourage buy-in and student investment, students were encouraged to take advantage of this "improvement policy," and to do so, they were required to turn in assignments on time and to show clear effort towards meeting the requirements. Conversely, while the control group was also told periodically that scientific writing was difficult and would take practice, no improvement options were given. That group was told only that all assignments must be completed on time or points would be deducted.

The write up assignments also included two other components. The first was a series of questions designed to help them think through the figure. Students brought the questions with them to their lab sections where they were encouraged to talk with their group mates and with their Teaching Assistants to get feedback on their thought processes. This was due at the start of their lab section. The write up was due the following day at midnight to allow them time to incorporate any feedback. The last component, which was turned in along with the write up, was a series of metacognition questions that asked students to analyze the tools they used to complete the write up, to compare levels of difficulty in interpreting different types of data figures, and what aspects of their write up they attempted to improve.

Grading of the anonymized assignments was conducted by undergraduate and graduate teaching assistants. Teaching assistants graded the first assignment alongside the instructor after a "rubric norming" session to ensure consistent grading. The second two assignments were graded independently by the teaching assistants. Often, the same teaching assistant graded the same group of students each time; however, given time constraints and the fact that assignments were anonymized and could only be classified

by lab section, this was not possible in all cases. In some instances, the same student may have had each of their write ups graded by a different teaching assistant.

A researcher-developed survey was created and distributed to students through the course's online platform at the end of the semester, in which students who completed it and took a screenshot of their completed screen as part of a regular assignment for the course. The survey included 21 questions which addressed students' perceptions of improvement in critical thinking skills, writing skills, content knowledge, as well as their perceptions of moving forward with STEM topics/learning as a result of participating in the course (see Appendix B). A Cronbach's Alpha coefficient of 0.922 was computed for the survey items, indicating that the survey was a reliable instrument to measure students' perceptions. Two qualitative questions also were provided for students' feedback.

All communication regarding the improvement policy was delivered via the online platform to the randomly assigned groups in order to keep Teaching Assistants and the Instructor (also a principal investigator) blind to the students assigned to each treatment group. After the first writing assignment, constraints were implemented on the online platform to ensure that students saw treatment-group-specific feedback before being able to access the second and third writing assignments.

2.2. Data Analysis

Before analysis, data were assessed and recoded to combine the gender categories of women and those who chose nonbinary sexual identification, as the frequencies for nonbinary were small (< 3%). There was a total of 41 women/nonbinary persons (52.6%) and 37 men (47.4%) included in the analysis. Of these, 34 were in the control group (41.5%), and 48 were in the treatment group (58.5%). Within the control group, 56.2% (n = 18) were women/nonbinary persons, and 43.8% (n = 14) were men. Within the treatment group, 50% (n = 23) were women/nonbinary persons, and 50% (n = 23) were men.

Factor analysis using a principal components analysis (PCA) approach and a varimax rotation was conducted via SPSS Software (v28.0.1.0; IBM Corp, 2021) to determine what, if any, underlying structures existed for measures on the 17 survey items. PCA is an advantageous approach to factor analysis when the goal is to reduce the number of used independent variables when conducting multivariate techniques (Mertler & Vannatta, 2010). PCA was used to identify components, each of which were composed of a linear combination of correlated variables (i.e. the survey items), that explained a large proportion of the total variance across students in survey responses (Chumney, 2012). The identified components were each represented by a single computed eigenvalue representing the total variance explained by all survey items contained within a component. A varimax rotation was chosen due to its simplicity in identifying a small

number of components with large percentages of variance explained (i.e. large loadings) for a number of items (Abdi, 2003).

PCA produced a four-component solution evaluated as appropriate through criteria including eigenvalues, variance, scree plot, and residuals. Components, each representing multiple survey items, were then used as independent variables in subsequent analyses. This was done by transforming the individual data items loaded onto each component using eigen vectors to represent projections of data regarding eigen vector direction (Dutt, 2021). The Keiser-Meyer-Olkin (KMO), a test for measuring adequacy of sample size and adequacy for factor analysis (Shrestha, 2021), was not significant, indicating that data were adequate for factor analysis. Bartlett's Test of Sphericity, which tests that the variables are orthogonal, was significant ($p < .001$), indicating that the variables were related and independent and, thus, suitable for detecting underlying structures.

Cronbach's alpha coefficients were then computed for each component to measure their reliability (see Table 1 for components, items, alpha coefficients, and variance accounted for). The four-component solution was reduced to a three-component solution due to a low Cronbach's alpha computed for the fourth component (< 0.4), which only accounted for 9.1% of the variance in the items and was constructed of only two survey items, which were the following: 1. Compared to before this course, how likely are you to now continue in a STEM field, and 2. I enjoy writing about unfamiliar topics in STEM more as a result of the feedback I got this semester. This component was originally named Enjoyment of STEM before being removed from analysis (see Table 1).

Table 1. Reliability Statistics for the Created Components

Construct	Cronbach's alpha	% Variance Explained
Perceptions of improvement in critical thinking ● Writing about an unfamiliar topic ● Interpreting unfamiliar data in the media/public ● Interpreting unfamiliar data in Biology ● Applying critical thinking skills to an unfamiliar topic ● Writing about an unfamiliar topic ● I feel like my writing skills have improved as a result of the feedback I got this semester ● I feel like my critical thinking skills have improved as a result of the feedback I got this semester	.918	30.4%
Perceptions of improvement in content knowledge ● Improvement in ability to describe the structure and function of fundamental cell, tissue, and organ types in plants ● Improvement in ability to explain major events during the evolution of land plants ● Improvement in ability to describe land plants' diversity ● Improvement in understanding of plants' many uses and modifications	.868	16.2%
STEM motivation ● Writing is a skill which I can improve with hard work ● I enjoy thinking critically about STEM topics ● Interpreting scientific data is a skill which I can improve with hard work and practice ● Likeliness to continue in a stem field	.760	15.5%

After rotation, the first component was named Perceptions of improvement in critical thinking and accounted for 30.4% of the total variance in the original variables and included perceptions of improvement in the following: Writing about an unfamiliar topic, Interpreting unfamiliar data in the media/public, Interpreting unfamiliar data in Biology, Applying critical thinking skills to an unfamiliar topic, Writing about an unfamiliar topic, and Applying critical thinking skills to an unfamiliar topic, in addition to responses on the statements, I feel like my writing skills have improved as a result of the feedback I got this semester, and I feel like my critical thinking skills have improved as a result of the feedback I got this semester. The second component was named Perceptions of improvement in content knowledge and accounted for 16.2% of the variance in all variables. Four variables loaded on this component, including perceptions of improvement in ability to Describe the structure and function of fundamental cell, tissue, and organ types in plants; Explain major events during the evolution of land plants; Describe land plants' diversity; and Understanding of plants' many uses and

modifications. The third component accounted for 15.5% of the variance and was named STEM Motivation. The four variables which loaded on this component included Writing is a skill which I can improve with hard work, I enjoy thinking critically about STEM topics, Interpreting scientific data is a skill which I can improve with hard work and practice, and Likeliness to continue in a STEM field. Table 2 provides descriptive statistics for each of the three components for the original sample before removing outliers or those who preferred not to disclose their gender. Each component is measured as a scale variable around zero, meaning that higher numbers indicate higher levels of the measured variables, and thus, a greater frequency of sentiment trending towards the affirmative on the Likert scale.

Table 2. Descriptive Statistics

Descriptive Statistics

	Treatment vs. control	Mean	Std. Deviation	N
Perceptions of improvement in critical thinking	Control	-.2098754	1.03174337	33
	Treatment	.1800473	.93377099	48
	Total	.0211899	.98753272	81
Perceptions of improvement in content knowledge	Control	.0730732	1.08189823	33
	Treatment	-.0247464	.94352334	48
	Total	.0151060	.99677196	81
STEM motivation	Control	-.0238971	1.13365084	33
	Treatment	.0044570	.91703381	48
	Total	-.0070947	1.00415185	81

The researchers prepared to conduct a one-way multivariate analysis of covariance (MANCOVA) on the created components to assess whether significant differences existed across the three components and across the four gender and treatment combinations. MANCOVA is appropriate for examining differences in multiple continuous dependent variables across more than two groups, and works by further reducing the three dependent variables (components) to a single dependent variable combination to assess overarching trends aligning with gender, treatment, or their interaction. Before conducting the analysis, normality and linearity of the components across both gender and treatment group were assessed. All assumptions for a MANCOVA were met regarding the dependent variables, including linearity, which was assessed through normal Q-Q plots; normality, which was examined through histograms; and the absence of multicollinearity, which was assessed through calculating correlation coefficients. While six outliers were identified among the student respondents, we chose to leave these

within the dataset due to their being true outliers which represented legitimate observations of the studied population (Frost, 2019). For instance, outliers could be due to student responses on the Likert scale that differed from the majority of the class, but that included justification for their sentiment in their qualitative responses.

As all assumptions were met, the researchers then conducted a MANCOVA to determine whether differences existed in the three created components that served as the dependent variables—participants' perceptions of improvement in critical thinking, perceptions of improvement in content knowledge, and STEM motivation—based on whether the students were in the treatment or control group, and before and after controlling for gender. There were no outliers or multicollinearity among the three dependent variables. Box's Test of Equality of Covariance Matrices was not significant, meaning that there were no significant differences between the covariance matrices across the three components and the assumption of equality of covariance matrices was not violated, making Wilk's lambda an appropriate statistic to interpret (Mertler & Vannatta, 2010). All assumptions, including the equality of error variances and linearity, were met.

Analysis of covariance (ANCOVA) was also computed as a post hoc analysis for the three components individually as dependent variables. ANCOVA results were compared once it was determined that the interaction and main effects between gender and treatment were significant, for the three components combined, through the MANCOVA (see Table 3). This was done to assess where significant differences existed between gender and treatment groups for each individual component. A Bonferroni adjustment was made to counteract any potential for an inflated error rate which can occur from conducting multiple ANCOVAs (Mertler & Vannatta, 2010). PCA, MANCOVA, and ANCOVA were conducted using SPSS software (v28.0.1.0; IBM Corp, 2021).

Table 3. MANCOVA Summary Statistics

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.035	.860[b]	3.000	72.000	.466	.035
	Wilks' Lambda	.965	.860[b]	3.000	72.000	.466	.035
	Hotelling's Trace	.036	.860[b]	3.000	72.000	.466	.035
	Roy's Largest Root	.036	.860[b]	3.000	72.000	.466	.035
Group	Pillai's Trace	.101	2.701[b]	3.000	72.000	.052	.101
	Wilks' Lambda	.899	***2.701[b]***	3.000	72.000	***.052***	***.101***
	Hotelling's	.113	2.701[b]	3.000	72.000	.052	.101

		Trace						
		Roy's Largest Root	.113	2.701[b]	3.000	72.000	.052	.101
Gender1		Pillai's Trace	.089	2.358[b]	3.000	72.000	.079	.089
		Wilks' Lambda	.911	***2.358[b]***	3.000	72.000	***.079***	***.089***
		Hotelling's Trace	.098	2.358[b]	3.000	72.000	.079	.089
		Roy's Largest Root	.098	2.358[b]	3.000	72.000	.079	.089
Group * Gender1		Pillai's Trace	.102	2.722[b]	3.000	72.000	.051	.102
		Wilks' Lambda	.898	***2.722[b]***	3.000	72.000	***.051***	***.102***
		Hotelling's Trace	.113	2.722[b]	3.000	72.000	.051	.102
		Roy's Largest Root	.113	2.722[b]	3.000	72.000	.051	.102

*Note: Pillai's Trace, Wilk's Lambda, Hotelling's Trace, and Roy's Largest Root are all test statistics for the above MANCOVA and are provided for accuracy in reporting statistical values. Wilk's Lambda was used in this paper as Box's Test of Equality of Covariances was met. Pillai's Trace, Wilk's Lambda, and Hotelling's Trace provide more conservative statistics.

In addition to assessing overall differences across gender and treatment on the numerical responses to the survey questions, sentiment analysis was conducted to quantify students' written comments within the survey. Sentiment analysis, a computational examination of participants' sentiments or attitudes towards some studied topic (Medhat et al., 2014), can be helpful when analyzing qualitative data objectively. Sentiment analysis was conducted using R Software, Version 2022.12.0+353 (RStudio Team, 2020), to determine what underlying sentiments were present in participants' qualitative responses to the questions pertaining to their major takeaways from both the course and the feedback they received on their write-ups. Packages used included tidytext (v0.1.4; (Queiroz et al., 2023)); tm (0.7-11; (Feinerer, 2023)); lattice (v0.21-8; (Sarkar, 2008)); textdata (v0.4.4; (Hvitfeldt & Silge, 2022)); and scales (v1.2.1; (Wickham & Seidel, 2022)). Additionally, code was computed from a pre-existing script (Showrav, 2023). Examined sentiments were classified as either positive or negative, as defined by the aforementioned packages. After identifying sentiments, frequency plots were generated to visually analyze sentiment frequencies and trends for each group.

3. Results

MANCOVA was used to assess whether there were overall differences in survey response patterns among gender or treatment group across the three created components—Perceptions of improvement in critical thinking, Perceptions of improvement in content knowledge, and STEM motivation—combined into a single dependent variable. MANCOVA results are provided in Table 3. Results indicated that the interaction between group and gender was significant at the $p < .1$ level and almost significant at the $p < .05$ level (Wilk's Lambda = 0.102, [$F(1, 1) = 2.772$, $p = 0.51$, partial $\eta^2 = .102$]. The interaction between treatment group and gender accounted for approximately 10% of the variance in the combined dependent variable, indicating that at least a portion of students' perceptions, as measured by numerical survey responses, could be attributable to treatment group or gender.

Because of the context of the current study, we also interpreted the main effects of gender and treatment group on the combined dependent variable; however, these should be used for descriptive purposes only as the interaction was significant. Gender resulted in significant differences within the combined dependent variable at the $p < 0.1$ level (Wilk's Lambda = 0.911, [$F(1, 1) = 2.358$, $p = 0.079$, partial $\eta^2 = .089$]), although it accounted for less than 1% of the variance in the combined dependent variable. Treatment group resulted in significant differences within the combined dependent variable at the $p < 0.1$ level and almost at the $p < .05$ level (Wilk's Lambda = 0.899, [$F(1, 1) = 2.701$, $p = 0.52$, partial $\eta^2 = .101$], with group accounting for approximately 10% of the variance in the outcome variable. As treatment group accounted for a greater proportion of the variation, it is likely that grading style, rather than gender, accounted for the majority of the variation explained by the interaction.

Because the interaction effect and main effects on the combined dependent variable were significant at an alpha of 0.1, post hoc analyses were run through ANCOVAs to examine whether significant differences existed across gender and group on each of the three tested components, individually, as dependent variables. Each dependent variable was tested through the Bonferroni adjustment at a critical value of 0.017, which equaled the overall alpha of 0.05 divided by the number of dependent variables. The ANCOVA results of each component are discussed in turn.

For the component named Perceptions of improvement in critical thinking, the interaction between gender and treatment group was not significant, nor was the main effect of gender. Treatment group exhibited an almost significant difference ([$F(1, 1) = 5.092$, $p = 0.025$, partial $\eta^2 = .0066$]; see Table 4), although this was small, with those in the treatment group reporting 0.012 units (less than 1%) greater improvement in critical thinking skills than those in the control group. For the component named Perceptions of improvement in content knowledge¸ the interaction between gender and treatment group was again not significant, nor was the main effect of treatment group. However, there

were significant differences across gender ([F(1, 1) = 5.696, p < .017, partial η² = .077], with men perceiving significantly higher improvements in content knowledge compared to women/nonbinary persons. Finally, for the component named STEM motivation, there was a significant interaction effect between group and gender ([F(1, 1) = 6.468, p < .017, partial η² = .080]). Because there was an interaction effect on STEM motivation (see Figure 1), the main effects of group and gender were not interpreted. Specifically, while men reported approximately 15% higher STEM motivation in the control group compared to women/nonbinary persons, women/nonbinary persons reported approximately 8% higher STEM motivation compared to men in the treatment group. However, for each of the three components, the main effects of gender, treatment group, or their interaction, if significant, explained less than 1% of the overall variation observed across student survey responses. This indicates that while gender or grading style may be associated with differences in student perceptions, other factors may be attributed to their perceptions to a greater degree.

Figure 1

Two-Way Significant Interaction Effect Between Group and Gender

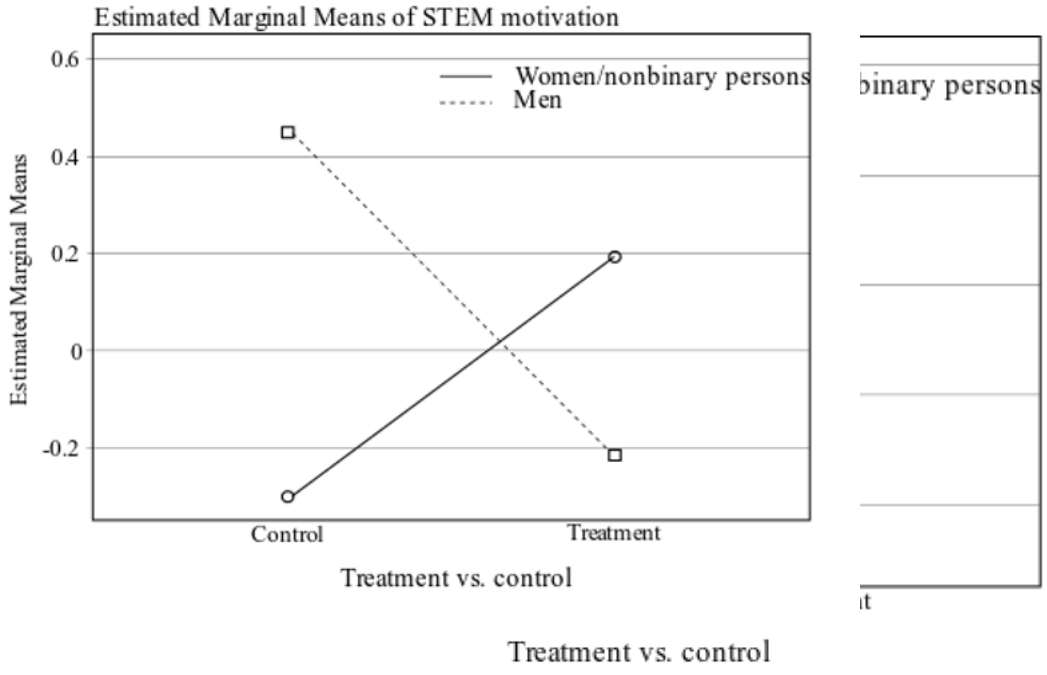


Table 4. Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	Perceptions of improvement in critical thinking	5.469[a]	3	1.823	1.88	.139	.071
	Perceptions of improvement in content knowledge	6.630[b]	3	2.210	2.38	.076	.088
	STEM motivation	6.319[c]	3	2.106	2.19	.096	.082
Intercept	Perceptions of improvement in critical thinking	.190	1	.190	.197	.659	.003
	Perceptions of improvement in content knowledge	2.107	1	2.107	2.27	.136	.030
	STEM motivation	.112	1	.112	.116	.734	.002
Group	Perceptions of improvement in critical thinking	5.092	1	5.092	***5.27***	***.025***	***.066***
	Perceptions of improvement in content knowledge	.512	1	.512	.552	.460	.007
	STEM motivation	2.434	1	2.434	2.53	.116	.033
Gender	Perceptions of improvement in critical thinking	.393	1	.393	.407	.526	.005
	Perceptions of improvement in content knowledge	5.696	1	5.696	***6.14***	***.015***	***.077***
	STEM motivation	.534	1	.534	.556	.458	.007
Group * Gender	Perceptions of improvement in critical thinking	1.651	1	1.651	1.70	.195	.023
	Perceptions of improvement in content knowledge	.149	1	.149	.161	.690	.002
	STEM motivation	6.218	1	6.218	***6.46***	***.013***	***.080***
Error	Perceptions of improvement in critical thinking	71.494	74	.966			
	Perceptions of improvement in content knowledge	68.651	74	.928			
	STEM motivation	71.140	74	.961			
Total	Perceptions of improvement in critical thinking	77.031	78				
	Perceptions of improvement in content knowledge	75.325	78				
	STEM motivation	77.461	78				

Adjusted and unadjusted group means for all participants across group and dependent variables, when controlling for gender, were also computed (see Table 5). Group means indicated that when gender was controlled for, those in the treatment group reported higher perceptions of improvement in critical thinking and STEM motivation compared to those in the control group, but lower perceptions of improvement in content knowledge.

Table 5. Descriptives

	Treatment vs. control			Statistic	Std. Error
Perceptions of improvement in critical thinking	Control	Mean		-.2541844[a]	.17978653
		95% Confidence Interval for Mean	Lower Bound	-.6199628	
			Upper Bound	.1115941	
		Std. Deviation		1.04832658	
	Treatment	Mean		.1800473[a]	.13477823
		95% Confidence Interval for Mean	Lower Bound	-.0910916	
			Upper Bound	.4511861	
		Std. Deviation		.93377099	
Perceptions of improvement in content knowledge	Control	Mean		.0349361	.18664884
		95% Confidence Interval for Mean	Lower Bound	-.3448038	
			Upper Bound	.4146760	
		Std. Deviation		1.08834043	
	Treatment	Mean		-.0247464	.13618586
		95% Confidence Interval for Mean	Lower Bound	-.2987170	
			Upper Bound	.2492242	
		Std. Deviation		.94352334	
STEM motivation	Control	Mean		-.0062922	.19225884
		95% Confidence Interval for Mean	Lower Bound	-.3974457	
			Upper Bound	.3848614	
		Std. Deviation		1.12105203	
	Treatment	Mean		.0044570	.13236243
		95% Confidence Interval for Mean	Lower Bound	-.2618219	
			Upper Bound	.2707358	
		Std. Deviation		.91703381	

Subscript[a] represents a significant difference in means.

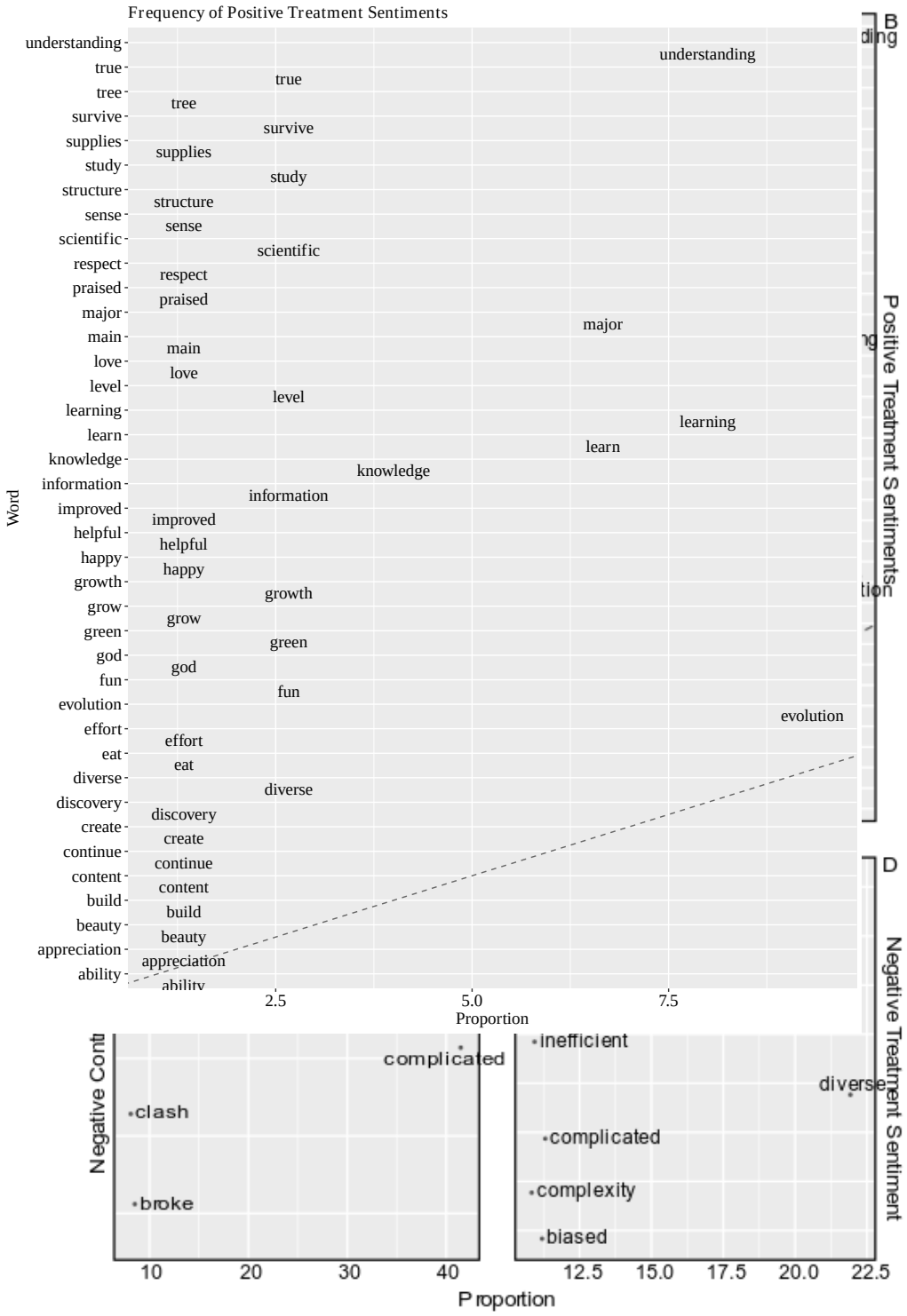
3.1. Sentiment Analysis Results

To better understand the potential differences in student perceptions based on whether they received a growth-based grading style compared to a traditional grading style, a sentiment analysis was conducted to identify proportions of words which accounted for positive or negative sentiments in response to a qualitative question regarding students' major takeaways from the course and feedback received on their write ups between groups, regardless of their gender (see Figure 2). Sentiment analysis allowed the researchers to better understand the words that participants used to describe their experiences, whether positive or negative, across treatment groups. Overall, there was a lower proportion of words pertaining to negative sentiments (see Figure 2) than positive sentiments (see Figure 2) in both treatment groups.

There were many positive sentiments that overlapped in both groups, including references to content, enjoyment, and learning. Those in the treatment group referred positively to their experience of growth, while those in the control group did not. Although there was a lower proportion of negative sentiments overall across the treatment groups, between groups the control group expressed fewer negative sentiments than the treatment group, and the negative sentiments differed between treatment groups. While the control group had negative sentiments regarding the subject, including it being complicated and diverse, those in the treatment group used words like "misleading," "opposed," and "biased." Both groups referred to the writing assignments as being diverse and complicated. Thus, overall, students appeared to view the course and writing assignments more positively than negatively, although the positive and negative sentiments differed across treatment groups.

Figure 2

Negative Sentiments for Treatment Group: Proportion of Word Frequencies



4. Conclusions and Discussion

The aim of this work was to assess whether receiving growth-based grading (the treatment), as opposed to traditional grading (the control), correlated with differences in students' perceptions of course outcomes in a General Botany course, and whether these perceptions were altered based on the gender of the individual. To assess this, the researchers compared students' quantitative (Likert scale) and qualitative responses to a 21-question survey. The quantitative survey responses were correlated so as to allow reduction to three component variables: Perceptions of improvement in critical thinking, Perceptions of improvement in content knowledge, and STEM motivation, and analysis using a one-way multivariate analysis of covariance (MANCOVA) of the three component variables combined. MANCOVA results indicated that the factor interaction of treatment group and gender was significant and accounted for approximately 10% of the explained variance in the combined dependent variables, implying that the interaction between gender and treatment group resulted in significant differences across students' STEM motivation, perceptions of improvement in content knowledge, and perceptions of improvement in critical thinking skills. Regarding main effects, which were also interpreted for the purposes of the current study, gender resulted in significant differences within the combined dependent variable yet only accounted for 1% of the explained variance. There were also significant differences by treatment group in the combined dependent variable, accounting for 10% of the explained variance. Again, these main effects should be interpreted with caution; however, this may indicate that the majority of explained variance in the combined dependent variable stemmed from treatment group, rather than gender.

As the interaction was significant, analysis of covariance (ANCOVA) results were interpreted for each of three component variables, individually. Here, it was noted that there were significant differences across gender, with men perceiving higher Perceptions of improvement in content knowledge compared to women/nonbinary persons, while neither treatment group nor the gender-by-treatment group interaction were significant for that component variable. This is consistent with previous results that show that perceptions of men relative to women in STEM achievement do differ. For instance, men consistently selected other men as being knowledgeable about course content, even when controlling for actual course achievement and the degree of outspokenness of the individuals, and they overestimated the grades of male peers, whereas women tended to be more gender-neutral in their selections (Grunspan et al., 2016). Further, cultural perceptions of gender norms as applied to individuals in STEM can influence perceptions of success, specifically as it relates to women in STEM (Selimbegović et al., 2019). Whether the results of this current study reflect accurate perceptions of improvement in the General Botany course or reflect students' perceptions of gender norms in STEM remains to be determined. Additionally, it has been shown that individuals within the

LGBTQ+ community, including gender non-conforming individuals, do face exclusionary behavior in STEM (Patridge et al., 2014), which could affect sense of belongingness and shared identity with the course field of study. However, such trends cannot be confirmed here, as the number of individuals identifying as non-binary were small.

For the component variable called STEM motivation, there was a significant interaction between treatment groups and gender, in which men reported higher motivation in the control group, while women reported higher motivation in the treatment group. Previous work has demonstrated that highly-structured, active-learning environments that emphasize higher-order cognition improves the performance of all students but has a disproportionate advantage for students from underrepresented groups (Haak et al., 2011). The structure used here, in which students answered questions about the data figures and then discussed them in groups and with their teaching assistants before completing their write ups, is consistent with this approach. Later work has demonstrated that a likely reason for this strong benefit is because this high level of structure coupled with higher-level cognitive practice reduces barriers such as stereotype threat (Jordt et al., 2017). While this component variable assesses motivation rather than ability, this difference in perception among women in the treatment groups may indicate that growth-mindset interventions reduce perceived cultural barriers to STEM careers (Piatek-Jimenez et al., 2018; Starr & Simpkins, 2021; Lapytskaia Aidy et al., 2021), especially in the rural south where gender norms are still highly defined (Huffmon et al., 2016; Kamke et al., 2022). The interaction in which men report lower STEM motivation in the treatment group compared to women is consistent with a previous study in which women who endorsed a growth mindset increased their performance expectations in math and also achieved higher math grades than men (Degol et al., 2018). The authors attributed this difference to greater sensitivity of women to the detriments of a fixed mindset, perhaps due to the effects of cultural norms. In our current work, while STEM motivation declined among men in the treatment group compared to men in the control group, motivation increased in women relative to men in the treatment group. Thus, it may be that women benefit more strongly than men from growth-mindset interventions.

While there were significant differences in two of the three component variables, they explained only a minor portion (1%) of the total variance across student survey responses. Other factors, not tested in this work, likely explain more of the variation. It is possible growth-based grading plays a more pronounced role in subsets of students who were not factored into this study. For instance and in one past study, although all students benefitted from an active-learning environment in which critical thinking was emphasized, students from underrepresented groups, including first-generation students, benefitted more than students from groups most represented in STEM (Haak et al., 2011; Jordt et al., 2017). Although the majority of our students were White, and thus, we could not account for the effect of race, other structural descriptors, such as students from rural

high schools or who are the first in their family to attend college, could explain more of the variation observed. These factors were not accounted for in this work.

Finally, to better understand the perceptions of students as they related specifically to the writing assignments for which growth-based grading was used, the researchers conducted a sentiment analysis on students' qualitative survey responses. Sentiment analysis results indicated higher frequencies of positive sentiments towards the course and feedback received on write ups than there were negative sentiments in both groups. This indicates that students, overall, valued the assignment which was designed to build higher-order cognitive skills of evaluation and synthesis and to teach and reinforce the underlying structure of scientific papers and presentations—all skills with broad relevance. Consistent with this observation, in an analysis of 46 4-year institutions in the U.S., academic rigor was associated with greater self-motivation in learning, especially for students who entered college with low ACT scores or less positive attitudes about literacy (Culver et al., 2019). Additionally, academic rigor in class, as opposed to rigorous exams, disproportionally benefitted first-generation students. Similarly, an inquiry-based bioinformatics workshop that incorporated student writing of condensed scientific articles—in line with what was conducted here—correlated with greater student engagement than a workshop that did not include the writing assignments, despite initial hesitancy to write (Jeon et al., 2021). This initial hesitancy and resistance that abates through the course is consistent with anecdotal observations by the teaching assistants and instructors here. These results indicate that rigorous assignments, especially those with in-class components, may boost student motivation. Additionally, short, but realistic, writing assignments appear to be a useful vehicle by which to build higher-level cognitive skills into STEM classrooms.

Consistent with the goal of this study, those in the treatment group referred positively to the growth they experienced, while those in the control group did not. This reinforces the idea that instructor interventions can strengthen or lead to the acquisition of a growth-based mindset in students. This result is consistent with the outcomes of interventions intended to demonstrate that failures are normal and that failures can be overcome, and the outcomes of such interventions can be long lasting (Brady et al., 2020). However, those in the control group recorded fewer negative sentiments than those in the treatment group. Specifically, those in the treatment group expressed words like, "misleading", "opposed", and "biased". One possible reason for these differences in negative sentiments is that on the assignments for which growth-based grading was used, students were asked to learn from the feedback given in previous assignments to improve their overall grades. It is likely that some feedback appeared confusing or that grading appeared inconsistent from one assignment to the next. This may reflect the slightly different expectations of the instructor relative to the teaching assistants, despite the effort to normalize grading expectations based on the rubric. This may also reflect the fact that student work was not necessarily graded by the same grader each time or that

most feedback was provided through the electronic interface so as to keep the principal investigator blind to the treatment groups to which each student belonged. However, these sentiments also imply that students in the treatment group were actively reading and responding to grader feedback, perhaps more so than those in the control group, although both received the same feedback. In sum, both the positive and negative sentiments imply that students in the treatment group internalized, at least to some degree, the growth mindset, and then actively worked to improve writing from one assignment to the next.

5. Limitations & Recommendations for Future Research

While this study indicates the likely positive association between incorporating academically rigorous assignments into a STEM classroom when coupled with growth-mindset interventions and student perceptions, there were a few limitations. First, the high number of students and the need to maintain separate types of communications between the two treatment groups necessitated grading by teaching assistants and electronic communication of feedback through an online platform. This likely resulted in grading inconsistencies and a lack of clarity that may have been minimized had the instructor been able to speak freely. However, despite some negative sentiments apparently associated with grading, student sentiments were overall positive, indicating the viability of our approach. Future studies may benefit from more time for "rubric norming" between the instructor and the teaching assistants or group grading for all assignments rather than just the first. Second, these data included participants' perceptions of growth in critical thinking skills and content knowledge, rather than actual assessment of student ability. Future studies should assess whether differences in grading methods result in significant differences in students' actual writing, critical thinking skills, and content knowledge. Future studies may also benefit from assessing the influence of additional structural features describing the students surveyed, such as socioeconomic background, size and resources of their high schools, initial ability, or whether students were the first in their families to attend college. Because the sample was predominantly White and race was not assessed, future studies should also assess whether race plays a role in the relationship between growth-based grading and student outcomes. Furthermore, because this study used a quasi-experimental design and groups were randomly sampled from a larger convenient sample, any correlations cannot infer causation. Therefore, it is important to note that other confounding variables may be at play in the relationship between type of grading environment and students' perceptions.

References

Abdi, H. (2003). Factor rotations in factor analyses. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds.), *Encyclopedia of Social Sciences Research Methods*. Sage.

Brady, S. T., Cohen, G. L., Jarvis, S. N., & Walton, G. M. (2020). A brief social-belonging intervention in college improves adult outcomes for black Americans. *Science Advances*, *6*(18). https://doi.org/10.1126/sciadv.aay3689

Campbell, C. M., Smith, M., Dugan, J. P., & Komives, S. R. (2012). Mentors and college student leadership outcomes: The importance of position and process. *The Review of Higher Education*, *35*(4), 595–625. https://doi.org/10.1353/rhe.2012.0037

Casad, B. J., & Bryant, W. J. (2016). Addressing stereotype threat is critical to diversity and inclusion in organizational psychology. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00008

Cheryan, S., Meltzoff, A. N., & Kim, S. (2011). Classrooms matter: The design of virtual classrooms influences gender disparities in computer science classes. *Computers & Education*, *57*(2), 1825–1835. https://doi.org/10.1016/j.compedu.2011.02.004

Chumney, F. (2012). Principal components analysis, exploratory factor analysis, and confirmatory factor analysis. *West Georgia University*. https://www.westga.edu/academics/research/vrc/assets/docs/PCA-EFA-CFA_EssayChumney_09282012.pdf

Culver, K. C., Braxton, J., & Pascarella, E. (2019). Does teaching rigorously really enhance undergraduates' intellectual development? The relationship of academic rigor with critical thinking skills and lifelong learning motivations. *Higher Education*, *78*(4), 611–627. https://doi.org/10.1007/s10734-019-00361-z

Davies, P. G., Spencer, S. J., & Steele, C. M. (2005). Clearing the air: Identity safety moderates the effects of stereotype threat on women's leadership aspirations. *Journal of Personality and Social Psychology*, *88*, 276–287. https://doi.org/10.1037/0022-3514.88.2.276

Dee, T. S. (2014). Stereotype threat and the student-athlete. *Economic Inquiry*, *52*(1), 173–182. https://doi.org/10.1111/ecin.12006

Degol, J. L., Wang, M.-T., Zhang, Y., & Allerton, J. (2018). Do growth mindsets in math benefit females? Identifying pathways between gender, mindset, and motivation. *Journal of Youth and Adolescence*, *47*(5), 976–990. https://doi.org/10.1007/s10964-017-0739-8

Dennehy, Tc., & Dasgupta, N. (2017). Female peer mentors early in college increase women's positive academic experiences and retention in engineering. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(23). https://doi.org/10.1073/pnas.1613117114

Dutt, A. (2021, October 18). *A step by step implementation of principal component analysis*. medium. https://towardsdatascience.com/a-step-by-step-implementation-of-principal-component-analysis-5520cc6cd598

Dweck, C. S., & Yeager, D. S. (2019). Mindsets: A view from two eras. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *14*(3), 481–496. https://doi.org/10.1177/1745691618804166

Feinerer, I. (2023). *Introduction to the tm package text mining in R*. https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Frost, J. (2019, October 23). *Guidelines for removing and handling outliers in data*. Statistics By Jim. https://statisticsbyjim.com/basics/remove-outliers/

Grunspan, D. Z., Eddy, S. L., Brownell, S. E., Wiggins, B. L., Crowe, A. J., & Goodreau, S. M. (2016). Males under-estimate academic performance of their female peers in undergraduate biology classrooms. *PLOS ONE*, *11*(2), e0148405. https://doi.org/10.1371/journal.pone.0148405

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, *332*(6034), 1213–1216. https://doi.org/10.1126/science.1204820

Hamrick, K., Rivers, E., Arora, V., Finamore, J., Aydin, M., & Adeshiyan, S. (2021). *Women, minorities, and persons with disabilities in science and engineering: 2021.* (Special Report NSF 21-321). National Science Foundation, National Center for Science and Engineering Statistics. https://ncses.nsf.gov/wmpd

Huffmon, S. H., Lawrence, C. N., & Briggs, A. (2016). Describing ourselves: Identity overlap and fault lines regarding how southerners would describe the south to non-southerners. *HJEAS: Hungarian Journal of English and American Studies*, *22*(1), 53-78,240,245-246,249-250.

Hvitfeldt, E., & Silge, J. (2022). *Textdata: Download and load various text datasets* (0.4.4) [Computer software]. https://cran.r-project.org/web/packages/textdata/index.html

Jeon, A.-J., Kellogg, D., Khan, M. A., & Tucker-Kellogg, G. (2021). Developing critical thinking in STEM education through inquiry-based writing in the laboratory classroom. *Biochemistry and Molecular Biology Education*, *49*(1), 140–150. https://doi.org/10.1002/bmb.21414

Jordt, H., Eddy, S. L., Brazil, R., Lau, I., Mann, C., Brownell, S. E., King, K., & Freeman, S. (2017). Values affirmation intervention reduces achievement gap between underrepresented minority and white students in introductory biology classes. *CBE—Life Sciences Education*, *16*(3), ar41. https://doi.org/10.1187/cbe.16-12-0351

Kamke, K., Widman, L., & Javidi, H. (2022). The multidimensionality of adolescent girls' gender attitudes. *Gender Issues*, *39*(2), 236–251. https://doi.org/10.1007/s12147-021-09288-1

Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin*, *137*(4), 616–642. https://doi.org/10.1037/a0023557

Lapytskaia Aidy, C., Steele, J. R., Williams, A., Lipman, C., Wong, O., & Mastragostino, E. (2021). Examining adolescent daughters' and their parents' academic-gender stereotypes: Predicting academic attitudes, ability, and STEM intentions. *Journal of Adolescence*, *93*, 90–104. https://doi.org/10.1016/j.adolescence.2021.09.010

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mertler, C. A., & Vannatta, R. A. (2010). *Advanced and multivariate statistical methods: Practical application and interpretation* (2nd ed.). Pyrczak Publishing.

Patridge, E. V., Barthelemy, R., & Rankin, S. R. (2014). Factors impacting the academic climate for LGBQ STEM faculty. *Journal of Women and Minorities in Science and Engineering*, *20*(1). https://doi.org/10.1615/JWomenMinorScienEng.2014007429

Piatek-Jimenez, K., Cribbs, J., & Gill, N. (2018). College students' perceptions of gender stereotypes: Making connections to the underrepresentation of women in STEM fields. *International Journal of Science Education*, *40*(12), 1432–1454. https://doi.org/10.1080/09500693.2018.1482027

Queiroz, G. D., Fay, C., Hvitfeldt, E., Keyes, O., Misra, K., Mastny, T., Erickson, J., Robinson, D., Silge [aut, J., & cre. (2023). *Tidytext: Text mining using "dplyr", "ggplot2", and other tidy tools* (0.4.1) [Computer software]. https://cran.r-project.org/web/packages/tidytext/index.html

RStudio Team. (2020). *RStudio: Integrated development environment for R*. http://www.rstudio.com/

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. Springer. https://doi.org/10.1007/978-0-387-75969-2

Selimbegović, L., Karabegović, M., Blažev, M., & Burušić, J. (2019). The independent contributions of gender stereotypes and gender identification in predicting primary school pupils' expectancies of success in STEM fields. *Psychology in the Schools*, *56*(10), 1614–1632. https://doi.org/10.1002/pits.22296

Showrav, R. (2023, January 22). Text mining and sentiment analysis in R. *MLearning.Ai*. https://medium.com/mlearning-ai/text-mining-and-sentiment-analysis-in-r-f6cc7944d540

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), Article 1. https://doi.org/10.12691/ajams-9-1-2

Starr, C. R., & Simpkins, S. D. (2021). High school students' math and science gender stereotypes: Relations with their STEM outcomes and socializers' stereotypes. *Social Psychology of Education*, *24*(1), 273–298. https://doi.org/10.1007/s11218-021-09611-4

Wickham, H., & Seidel, D. (2022). *Scale functions for visualization*. *RStudio*. https://scales.r-lib.org, https://github.com/r-lib/scales

Yeager, D. S., & Dweck, C. S. (2020). What can be learned from growth mindset controversies? *American Psychologist*, *75*, 1269–1284. https://doi.org/10.1037/amp0000794

APPENDICES

Appendix A. Rubric

	4	3	2	1 to 0
Appearance/Writing	Write up is neat and organized. Writing is clear and concise, grammar and punctuation are correct, and flow is logical. Required formatting is followed.	Write up is neat and organized. Writing is mostly clear, but not always concise. Grammar and punctuation are mostly correct, and flow is logical. Required formatting is followed.	Write up is neat and required formatting is followed. Flow is not clear and/or concise OR there are several grammar and/or punctuation errors.	Write up is not neat and organized. Flow is not clear and/or concise, AND there are several grammar and/or punctuation errors. AND/OR required formatting is not followed.
Title	Insightfully and accurately synthesizes findings/main conclusion into a cohesive, concise title.	Title accurately reflects the information in the figure, but is not concise OR does not describe findings/main conclusion (instead describes what was done, for example).	Title mostly accurately reflects the information in the figure, but is not concise AND does not describe findings/main conclusion (instead describes what was done, for example).	Title mostly does not accurately reflect the information in the figure.
Background	Provides sufficient information to enable reader to clearly understand question(s)/experiment(s) and their significance. Question(s) logically follow from information provided in background. No gaps in logic are apparent.	Provides sufficient information to enable reader to mostly understand the question(s)/experiment(s) and their significance. Question(s) mostly logically follow from information provided in background. There are only minimal gaps in logic.	Information is provided, but reader struggles to understand the significance of question(s)/experiment(s). Question(s) somewhat follow from information provided in background. There are several gaps in logic.	Background does not logically lead to the question(s)/experiments(s). Question(s) do(es) not follow from information provided in background.

Question	Question is clear, specific, suggests an experiment, and demonstrates a likely outcome. (I.e. POOR: how do different student-participation methods work in the classroom? GOOD: Does facilitated group discussion during the lecture period improve student comprehension over a traditional instructor-given lecture?).	Question is clear, specific, and suggests an experiment (i.e. lecture vs. group discussion) but may not include an outcome (i.e. comprehension). OR may include an outcome but doesn't suggest an experiment. OR portions of both are missing.	Question is clear and specific but doesn't suggest an experiment or outcome. OR Question has an experiment or outcome, but logic is not clear.	Question is not clear and/or specific.
Methods	Provides a clear description of procedures used, without adding unnecessary detail, such that methods are sufficient to understand stated observations. As applicable, description of treatment groups and controls are included. Reader could easily devise a similar experiment based on information provided. (Details like number of individuals per treatment are not necessary in this context.)	Provides a mostly clear description of procedures used such that methods are mostly useful for understanding the stated observations. Minor gaps in logic OR multiple unnecessary details included. Reader could devise a similar experiment based on information provided with effort.	Provides some description procedures used. Reader could not devise a similar experiment based on information provided OR methods are not sufficient to help reader follow stated observations.	Description of procedures used lacks several key components.

Observations	Observations are clearly described and are clear descriptions of the data (i.e. amounts, direction of change, etc.). Conclusions that can be drawn from the data are not included (i.e. "warm temperatures increased growth" is a conclusion and should NOT be included, whereas "20% more growth in the warm-temperature treatment" is an observation).	Observations are mostly clearly described and are mostly clear descriptions of what data show. May include some (minimal) conclusions OR requires effort to understand the main patterns in the data.	Observations include multiple conclusions OR data are mostly *not* clearly described, but attempt *is* made to describe what data show (amounts, directions of change, etc.). Reader could not understand patterns in data with information provided.	Several key details are missing OR this section includes only statements that would be considered conclusions (not descriptions the data).
Conclusion	Interpretation of the data logically follows from observations. Clarifies what the data actually show vs. what they suggest. Words like "prove" are avoided.	Interpretation of the data mostly logically follows from observations. Mostly clarifies what the data actually show vs. what they suggest, but some gaps may be present. Words like "prove" are typically avoided. OR Interpretation of the data is logical; however, statements such as "prove" are used inappropriately (i.e. conclusions overreach what the data actually show).	Interpretation of the data mostly do not logically follow from observations. Attempt is made to clarify what the data actually show vs. what they suggest, but there are clear errors. Words like "prove" may be included.	Conclusion does not include interpretation of observations. (for example, conclusion is a restatement of background or big-picture, or is a restatement of the observations). AND/OR words like "prove" are used.

Reflections on learning	Answers question or addresses prompt seriously and thoughtfully and provides specific examples from their own work to back up their points.	Answers question or addresses prompt seriously and thoughtfully. May provide examples from their own work to back up their points but they are not specific OR Provides some examples, but not all points are clearly explained.	Answers questions or prompt seriously but does not provide examples.	Does not show attempt to answer question or prompt seriously.

Appendix B. Survey

You are being asked to complete this survey because we are interested in your perceptions of what you learned this semester. **Please read each question carefully before responding.** The following survey should only take about 5 minutes to complete. Your responses will be kept confidential. Thank you for taking this survey.
The following questions ask about your confidence in and ability to work with data analysis, writing, and applying critical thinking skills as a result of taking this course.

Compared to **before** the semester and as a result of taking this course, my **CONFIDENCE** in ...

	About the same as before	A little better than before	A lot better than before
Writing about an unfamiliar topic is...	○	○	○
Interpreting unfamiliar data in the media/public is...	○	○	○
Interpreting unfamiliar data in Biology is...	○	○	○
Applying critical thinking skills to an unfamiliar topic is...	○	○	○

Compared to **before** the semester and as a result of taking this course, my **ABILITY** to...

	About the same as before	A little better than before	A lot better than before
Write about an unfamiliar topic is...	○	○	○
Apply critical thinking skills to an unfamiliar topic is...	○	○	○

The following questions ask about your perceptions of outcomes associated with your data analysis write-ups this semester.

Please indicate the extent to which you agree with the following statements about writing, as a result of working on data analysis write-ups throughout the semester.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I enjoy writing about unfamiliar topics in Science, Technology, Engineering, and Mathematics (STEM) more as a result of the feedback I got this semester.	○	○	○	○	○
I feel like my writing skills have improved as a result of the feedback I got this semester.	○	○	○	○	○

Writing is a skill which I can improve with hard work.	○	○	○	○	○

Please indicate the extent to which you agree with the following statements about data analysis and research, as a result of working on data analysis write-ups throughout the semester.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I enjoy thinking critically about STEM topics more than I did at the beginning of the semester.	○	○	○	○	○
I feel like my critical thinking skills have improved as a result of the feedback I got this semester.	○	○	○	○	○
Interpreting scientific data is a skill which I can improve with hard work and practice.	○	○	○	○	○

The following questions ask about your perceptions of course outcomes.
As a result of taking this course, my...

	About the same as before	A little better than before	A lot better than before
Ability to describe the structure and function of fundamental cell, tissue, and organ types in plants is...	○	○	○
Ability to explain major events during the evolution of land plants is...	○	○	○
Ability to describe land plants' diversity is...	○	○	○
Understanding of the importances of plants to survival of life on earth is...	○	○	○
Understanding of plants' many uses and modifications is...	○	○	○

The following questions ask about your general perceptions of the course and plans for the future.
What were your major takeaways from working on your data analysis write-ups this semester?

_____

_____

_____

What were your major takeaways from this course?

_____

_____

_____

How likely are you to continue in a STEM field after this semester?

- ○ Extremely unlikely
- ○ Somewhat unlikely
- ○ Neither likely nor unlikely
- ○ Somewhat likely
- ○ Extremely likely

Compared to **before** this course, how likely are you **now** to continue in a STEM field?

- ○ About the same as before
- ○ A little more likely
- ○ A lot more likely

The following questions ask about your gender and race. All responses will remain confidential, and the researchers will not be able to identify you with any information.

Please indicate the gender you most identify with socially.

- ○ Man
- ○ Woman
- ○ Non-binary / third gender
- ○ Prefer not to say