Quant II Final Paper

James Steur

05/15/2018

Abstract

This observational study examines the role of professors written comments on student papers to answer two questions: Do students actually improve their writing abilities from professors' comments? Do certain types of commenting encourage improvement in student writing? Two professors at a liberal arts college in the Rocky West provided their students' first and last papers from the spring of 2015 with their written comments on these papers after their classes were completed. First, researchers used a coding schema to identify all the different ways professors wrote comments on students' papers (n=382). Afterwards, researchers were blinded and evaluated the first and last papers' quality across four different outcome variables: global, sentence-level, local, and overall quality. Researchers conducted a paired sample t-test to see the difference in means between all outcome variables with all students pooled together and students separated by class. Researchers found statistically significant results for overall and macro outcomes when students were pooled together (p=0.001 and 0.005 respectively). When students were separated by class strata, researchers only found one statistically significant result for class 1's overall outcome variable (n=16, p=0.01) but not class 2 (n=10). Furthermore, professors commented in highly different ways: professor 1 mostly commented using evaluative methods, whereas professor 2 mostly commented with advising and directive comments. Although the sample size and power rates were small, these results could indicate that evaluative comments result in greater improvement. Results suggest that all students, regardless of the class or commenting style, improved their writing ability. Most students started roughly at a "C" level, and improved to a final grade of "C+" or "B-" by the end of the class. This suggests that both professors commenting styles together show improvement. There are two primary directions for future research: 1). Researchers using textual analysis or machine learning algorithms of professors comments to obtain larger n-sizes. 2). Researching different types of students, like first-generation or non-native English speakers, to see if certain types of comments help different types of student improve their writing ability more than other types of commenting.

Part I

Teachers Disagree About How to Teach Writing

How do people learn to write well?" A wide literature exists on the theory and teaching of writing to students. Expressivism stresses that teachers should cultivate writers' unique voices through creative exercises such as poetry and force the writer to make their own choices (Murphy & Sherwood, 1995). Current traditionalists emphasize that writing follows a "correct" set of identifiable rules, and writers must follow these rules to produce excellent papers. Interestingly, these and other theories lack one critical element: data. Teachers, professors, and researchers make different claims about different theories of writing that help students learn the most. However, proponents of these claims often ground their arguments in theory or logic. Ultimately, conversations about which theory educators should use, like current traditionalism or expressivism, ends in "I teach this way; you teach your way."

This "agree to disagree" mentality has led the literature to focus on students' perceptions of feedback on their papers (Sommers, 2013). For instance, students feel attacked and morally judged with "excessive" commenting on their papers. Conversely, students who receive "too few" comments desire additional feedback. ##Missing Gap: How should we comment? Why should we care?

At this point, almost no literature uses quantitative data to support specific theories and styles of teaching writing (Anson, 2008). Moreover, no research demonstrates that students improve their writing from a specific style of paper commenting. The lack of research on how professors' influence student learning is important for a variety of reasons. Teachers should care about research on writing pedagogy to discover what types of comments help students learn. Rather than teachers offering feedback on untested theories and unfounded data about what helps student learn, this line of research can offer specific commenting styles that can improve student learning outcomes. Indeed, college campuses generally assume that first year general education English courses are effective at teaching writing to college students, but there is little evidence supporting this point.

An extended implication of extending this research is reproducible research. At this point, the field of writing studies is plagued with anecdotal examples and theories that are untested with data. Researchers and educators anecdotally note improvements in their students' abilities, however, there is no existing database, record, or meta-analysis to support these claims. To some extent, educators institutions simply told them a way to teach writing, and they reteach their students the same way they were told. Educators often rely on their perceptions of student learning, which could be biased in their interpretation of student learning. Identifying specific commenting styles on student papers could encourage future college campuses and public schools to adopt commenting styles that are effective in helping students learn how to write.

In the end, professors can understand how their comments influence students' writing and thinking from conducting research, but there is a large hole in the literature. All of this existing literature, or lack thereof, drove my primary research question: How do professors' comments on first year English composition students influence student writing? At a broad level, I want to identify **any** improvement that commenting has on student writing. I also explore a related sub-question: Do certain styles of commenting result in more improvement?

Causality

This study primarily aims to see the effects of a cause rather than the causes of an effect. One reason for this emphasis is data limitations. Unfortunately, researchers only knew the participants were first-year students in their second semester of college—they had no other identifying information. Given this limitation, it is impossible to identify all the potential lurking varaibles that could influence student performance. Perhaps a student had a death in the family which impacts their school performance; perhaps some students are simply more motivated; perhaps some students came from high schools that prepared them better for college. In short, there are many lurking variables that confound the specific causal mechanisms of the design and its results. Consequently, this study focuses on showing the effects of a cause and discussing INUS conditions that could influence these results.

Data Collection

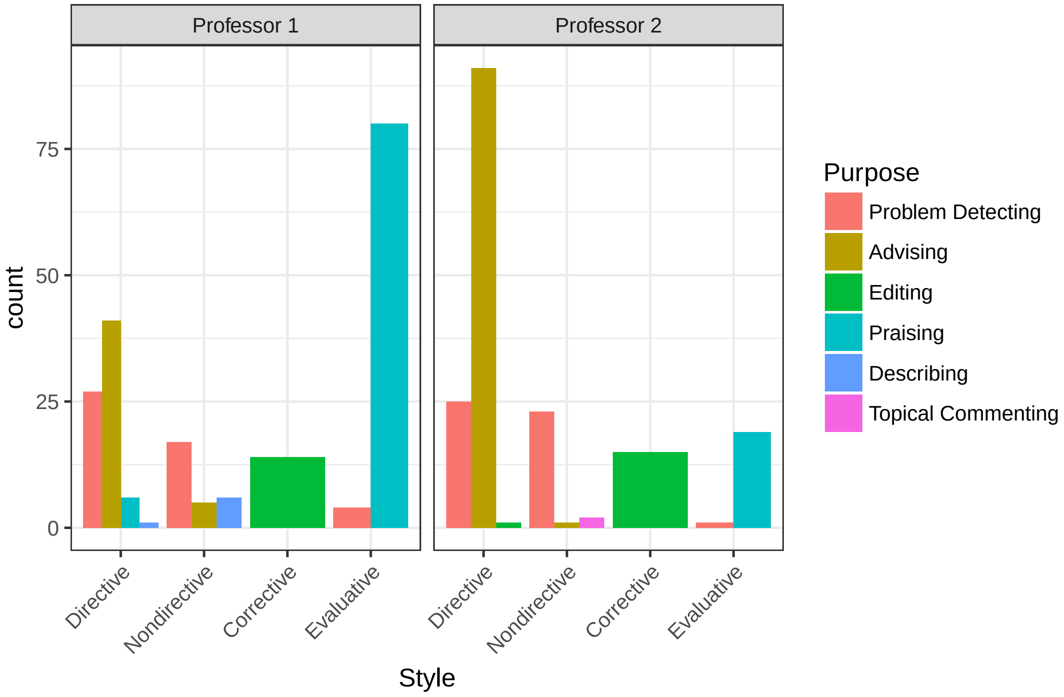
This observational study was conducted at a liberal arts college in the Rocky Mountains over the summer of 2015. An email was sent to all professors in May of 2015 who taught first-year English courses to assist with this research project, and two professors agreed to help. As a result, my project used a convenience sample of students. Afterwards, communication occurred between the researchers and professors in order to collect appropriate materials. Both professors agreed to provide their students' first and last papers for their classes in the spring of 2015. Researchers collected papers after professors finished teaching the class to reduce performative commenting. That is, researchers wanted to ensure professors commented on student papers in a natural setting without observation. If professors knew researchers would evaluate their comments, they may comment differently since they are being observed, which would influence the results of the study.

All participants names and grades were omitted from these papers for evaluation purposes. In addition, the researchers were blinded from any identifying information about the students to reduce bias in the results. Then, two researchers used a modified code scheme (Beason, 1993) to assess the different dimensions of professors' commenting styles on student papers across four different categories: purpose, style, level, and article (see Appendix 1 for a more thorough discussion). Researchers did not code end comments that included overall strengths and weaknesses of the paper. Researchers made this choice for one primary reason: professors summarized the bulk of their comments they made throughout the paper. As a result, researchers did not want to "double code" certain kinds of comments into the analysis and skew the data in one direction. After researchers coding all comments on the papers, n=382. The first class had 201 written comments and the second class had 178 written comments. Rather than describe all aspects of how professors commented, I looked at how professors stereotypically comment on student papers. Looking at stereotypical commenting patterns illuminate what feedback helps student learning.

Professor one's most frequent types of comments are praising, evaluative, and macro-oriented that discuss the student's logic and reasoning, organization, and use of evidence. Some examples of this type of commenting would include, "I think you did an excellent job organizing this paragraph. This chain of reasoning doesn't sound quite right to me. I like the evidence you introduce in this article." Professor two's most frequent types of comments are advising, directive, and macro-oriented comments. However, professor two's concentration of different macro articles is scattered throughout the coding schema of Appendix 1. The most frequent article that professor two addresses is about the works cited page and citation errors. Some examples of what professor two's commenting style looks like include, "This in-text citation is incorrect—you should put the period outside the parenthesis. In APA style, you should italicize the title of the journal. I think you could better organize this paragraph by shifting these sentences around."

I have created Figure 1 to highlight the stark differences in the purpose and style of each professors commenting. As a reminder, I and another researcher coded each comment a professor made across seven dimensions: class, paper, student, purpose, style, level, and article. By highlighting the frequency of purpose and style together, it is easier to visualize how each professor comments on student papers. Importantly, this figure combines all comments from the first and last paper of the semester for both professors. Both professors had distinct commenting styles and combining comments from the first and last paper highlight these differences.

Figure 1: Overall Professor Comments



As Figure 1 indicates, professor one has over 75 comments on student papers that are praising and evaluative, and professor two has over 87 comments on student papers that are advising and directive. Both ways of commenting on student papers are drastically different from one another.

Professors written comments are a critical element that influences student writing. Importantly, professors general commenting trends simplify the data, and there are more factors that play into student learning. However, highlighting the most common approaches that both professors take is useful in differentiating their approaches to commenting. Moreover, these comments are the specific causal pathway that most closely corresponds to student improvement in their writing. The only way for students to learn writing is to do it, and professor comments on student work is the primary mechanism that influences their improvement. [1]

Afterwards, two researchers used a rubric to assess the quality of all students' papers across four outcome variables: macro, mid, micro, and overall quality. The micro outcome variable grades the student's paper on his or her grammar, punctuation, spelling, and capitalization. The mid variable grades the student's paper on his or her clarity, wordiness, or lack of variety in sentence strucutre. The macro outcome variable grades the student's paper on his or her awareness of the audience, support of

[1] Interested readers can see contingency tables of all professors commenting purposes, styles, levels, and articles in Appendix 2.

claims with sufficient and appropriate examples, evidence from authoritative sources, and organzation. The overall variable grades the student's paper on his or her thesis statement, paragraph structure, transitions, syntax of sentences, grammar, and the overall paper with all aspects considered. I scaled all of these otucome variables from 1 (poor) to 4 (excellent). In this design, a score of 1 is considered a grade of D, 2 is a C, 3 is a B, and 4 is an A (see Appendix 3). Herein, I refer to all my outcome variables as overall, macro, mid, and micro for purposes of clarity.

Reviewers scored these different levels by reading the paper, deciding their scores, and sharing their scores with the fellow researcher. Thankfully, the researchers mostly gave the same scores for the different papers after sharing their analysis. In the few cases where the evaluators disagreed on a score, they returned to the rubric and agreed on a final score before moving forward. After sharing their respective scores, most of the scores for the paper scorers were the same. Researchers did this for all four papers in the two different class sections: class one had sixteen students and class two had ten students.

Table 1: Variable Overall's Means, Ranges, and Coverage Intervals

	Means	Ranges	50% Coverage Intervals
Professor 1 First Paper	2.125	1 to 3	2 to 3
Professor 1 Last Paper	2.625	2 to 4	2 to 3
Professor 2 First Paper	2	1 to 3	2 to 2
Professor 2 Last Paper	2.500	1 to 3	2 to3

First Paper n=16. Second Paper n=10.

Table 2: Variable Macro's Means, Ranges, and Coverage Intervals

	Means	Ranges	50% Coverage Intervals
Professor 1 First Paper	2.250	1 to 4	2 to 3
Professor 1 Last Paper	2.688	2 to 4	2 to 3
Professor 2 First Paper	2	1 to 3	2 to 2
Professor 2 Last Paper	2.500	1 to 3	2 to 3

First Paper n=16. Second Paper n=10.

Tables 1 & 2 summarize the outcome variables for overall and macro across the four different papers researchers evaluated. The point estimates for overall professor 1 paper one and two are 2.125 and 2.625 respectively. The point estimates for professor 2 paper one and two are 2 and 2.5 respectively. Analytically, these results suggest that all students in both classes started at a "C" level of writing based on their first paper and ended with a "C+" or "B-" with their last paper grade. The 50%

coverage intervals are 2–3 for all professors and their papers, which suggest most students started and ended the class between a “C”–“B” level. At a descriptive level, there appears to be improvement across overall and macro after the students read the teachers comments.

Group Comparisons

The population of interest in this study is all undergraduate students who take a general education English course. Ultimately, I want to learn if students taking an English class actually learn or improve their writing abilities. In order to detect improvement, this study used a within-subjects design by looking at the students' first and last papers of the semester. The first paper for all students functioned as a baseline condition for their papers' writing quality. This was the first writing assignment all students completed at the start of the semester, so no other comments influenced these first papers' grades.

There are three reasons why the comparisons I draw between the students and professors are valid. First, all students are in their second-semester of their first year at college. If my sample had a mix of students in their first, second, third, and fourth year this would confound my results. Students who are older and take more classes may be more mature or write better than younger students. Since all students are in the same class in their college career, this removes possible variation in my sample. Second, all students took a standard English course from the college. That is, the English courses in this sample were not for non-native English speakers or honors students. It would be an unfair comparsion if one class section was honors students and the other non-native English speakers. Thankfully, all students in both class sections were in similar English courses. Third, the type of writing assignments in both classes were similar to one another. Both professors' prompts for their first papers were thesis-driven and asked their students to support their arguments with well reasoned evidence in 3–4 pages. The professors prompts for their second papers were virtually the same as the first except the paper was expected to be between 6–8 pages. If the papers between professors were different types of assignments, the comparsions would be unfair and confound any improvement my analysis would suggest. For example, if one paper were an annotated bibliography and the other a thesis-driven paper, it would be nonsensical to compare improvement between the two of them.

Strengths & Weaknesses

There are four strenghts to this study. First, all researchers on this project had over 2+ years in grading and evaluating student writing As a result, evaluators offered expertise on student writing, the rubrics, and the evaluation process in a skillful way. Second, researchers were blinded from all identifying information about the students and professors as they evaluated the papers. This reduced the possibility of biased grades. Moreover, all evaluators used the same rubric and came to a consensus about the grade all papers deserved. If researchers relied on the professors grades they gave to students, there could be variability in how harshly different professors grade. Some professors may mostly give “A’s” and “B’s,” whereas some may mostly give “C’s.” Thankfully, all papers were evaluated with the same scale by the same evaluators to ensure homogeneous comparisons. Third, all students first papers serve as a pre-treatment comparsion. This paper was the first assignment that all students completed, so no previous professor comments confounded the initial grades of these papers.

This study is limited in four primary ways. First, the researchers of this study were unable to obtain additional information about the students and match on observed covariates. It is possible, for instance, that some students came from high schools that prepared them for college better. It could also be possible that some students are simply more motivated or have disparate writing abilities at the start of the course. There could be a slew of pre-existing differences among students in the different groups that confound these results. A future study could identify observable covariates of the students and use matching as well as sensitivity analysis to provide additional evidence that professors written comments improve students’ writing abilities.

Second, this study had a small sample size of n=26 for the first part of this analysis and n=16 for class 1 and n=10 for class 2. Once participants were divided into professor strata, the statistical power of these analyses were sharply low. Future researchers should be aware of the time consuming nature of this research. Five different researchers conducted this analysis over the course of one summer for a final sample size of 26 students.

Third, this research only looks at the first and last comments of the different professors’ papers. In a classroom setting, there are other paper assignments and other written feedback being offered to students. These four papers cannot capture the totality of all the different commenting styles and learning that happens in the classroom. A longitudinal study that looks across all the different comments and improvements in paper scores could help capture a more granular level of learning.

Last, given the observational nature of the data, it is impossible to entirely rule out lurking variables that could confound the results. Lurking variables that I failed to measure could influence these results.

Part II

Pooled Students Methodology & Results

At baseline, I calculated the overall means for all students' first papers. The evaluation of all students' papers included my four outcome variables: the micro, mid, macro, and overall quality. Afterwards, I calculated the means of the students' second papers and looked at the average difference in means between the two papers. In this study, the "treatment" is the written comments professors provide to their students. Since I am interested in the differences between the two group means, I used a paired-sample t-test with α=0.05 (AERD Statistics, 2018). I compared these differences in mean with all the outcome variables for the first and last paper that students submit. The larger the difference in means between the two groups, the more likely it is that students are learning how to write more effectively from their courses. Suppose, for example, a student got an overall score of 1 on his or her first paper and a score of 4 on his or her last paper. This would be a mean difference of 3, which suggests the student improved by three letter grades: starting at a D and improving to an A. The smaller the difference in means between the two groups, the more likely it is that students did not learn how to write effectively. If the mean difference was 0, this would suggest no improvement between the first and second paper. Given the small sample size of n=26, the paired sample t-test also increases statistical power. By using a "repeated measures" test that compares differences within subjects, I am less likely to falsely reject the null hypothesis.

There are five primary assumptions of the paired sample t-test. For purposes of clarity, I outline each assumption and explain how my design addresses them.

1). The observations are independent of one another. For this paper, I define observations as the differences between two different sets of values (Boneau, 1960). In my design, I use a sample of students and measure the quality of their papers before and after professors provide written comments to them. Then, I analyze the difference in means of the outcome variables using a paired sample t-test. Essentially, I argue that one student's overall score on a paper does not influence a different student's overall score on their paper. Students are not influencing their peers improvement between papers since they do not write comments for one another on each other's papers.

2). The independent variable(s) should have two groups that are related to one another or function as "matched pairs" (NCSS, 2018; Boneau, 1960). This simply means the same subjects are present in both groups. In my study, all students in both groups have been measured on the same dependent variable at two different times: at the start and end of their class.

3). The dependent variable is measured on an interval or ratio scale (Hsu &

Lachenbruch, 2014). My dependent variables, the macro, mid, micro, and overall quality of the papers are all interval scales. The intervals between each of the numbers represent real equal differences in different dimensions of the papers. Put differently, the intervals between the values from 1-4 are meaningful. Moreover, the differences between the different categories have a consistent meaning with my measurement tool (see Appendix 3).

4). The dependent variable should not include outliers (Hsu & Lachenbruch, 2014; Boeau, 1960). If the dependent variable has outliers, this affects the skew of the distribution, which influences the mean of the data. If the mean of the data is greatly affected, then the mean will be biased and affect statistical inference with p-values and effect sizes. In my data set, there are no outliers in the dependent variable. All students received similar scores to one another.

5). The variances of the dependent variable are assumed to be equal to one another (NCSS, 2018 ;AERD Statistics, 2018). In my study, the variances are equal since I am dealing with the same group of students, so this assumption is not violated. If it were violated, I would need to conduct a different t-test that considers unequal variances in the dependent variable.

For the first part of my analysis, I choose to combine all of the students together for two reasons. First, my primary research question asks if students show improvement in their writing at any level. That is, can researchers see *any* effect from professors' comments on student writing? Although the professors comment in different ways, they are fundamentally a similar type of treatment: comments on student writing. Hopefully, after all students are pooled together, they show some improvement in their writing ability. (It would be disheartening if students showed no improvement that I could measure with statistics.) Second, the two student groups are similar in a variety of ways. They are all first-year students, they are all taking an introductory English course, and they are all in their second semester of undergrad. The student samples should be largely homogeneous groups.

After conducting the two-tailed paired sample t-test, I fail to reject the null hypothesis for both micro and mid-level improvement in student writing. Mid and micro's respective *p*-values are 0.2 and 0.2 (Table 3). Although it is possible that I am committing a type II error and falsely reject the null hypothesis, it appears that there is no statistically significant relationship between micro and mid-level improvement for student papers. In practical terms, this means it is unlikely that professors comments on student papers are improving students' ability to write prolific sentences. Whatever improvement students' demonstrate in mid or micro-level comments are most likely due to chance.

For my macro and overall outcome variables, I reject the null hypothesis. The paired sample t-test found *p*-values of 0.001 and 0.005 for the overall and macro outcome variables respectively (Table 3). I could be committing a type I error and have two false positive results, but it appears there is a statistically significant relationship between overall and macro-level improvement for student papers. In practical terms, it is likely that professors' comments on student papers are improving their overall papers and organization. The improvements we see from students' overall and macro level aspects of their papers are most likely due to professors commenting on their papers. In the end, there is evidence that suggests teachers'comments on student writing results in learning.

Table 3: Pooled Classes Paired Sample T-Test

Level	95% CI	t-statistic	p-value	Mean Difference
Overall	0.21 to 0.79	3.610	0.001	0.500
Macro	0.15 to 0.77	3.090	0.005	0.460
Mid	0.13 to 0.52	1.220	0.232	0.190
Micro	0.13 to 0.52	1.220	0.232	0.190

DF=25. Two-tailed test.

The mean differences in Table 3 represent how much improvement occured between all students' first papers to their last ones in my sample. The rubric evaluators scaled everything from 1 to 4, so an increase of 1 means a full letter grade increase, 0.5 means half a letter grade increase, and 0 means no improvement or deterioration. For example, overall's mean difference is 0.5, which suggests students improved by 0.5 from the evaluation rubic (Appendix 3). In substantive terms, this means that most students improved the overall quality of their papers by half a letter grade. For macro, the mean difference is 0.46, which also suggests students improved by half a letter grade. Mid and micro both have mean differences of 0.19, which suggests they made minor improvements in grammar and sentence syntax, but did not improve in any meaningful way.

However, sampling variability makes it difficult to look at one sample and exactly identify my population parameter. Instead of trusting a point estimate, I created confidence intervals to look at an interval of possible values. If I repeatedly collected samples, I would expect that 95% of the samples produce confidence intervals that capture my true population parameter (and 5% would not). Importantly, I can never know if my sample is one of the 95% that worked, which is an element of uncertainity I can't overcome. There is always the possibility that my confidence interval fails to capture the true population parameter.

In the table above, I presented four 95% confidence intervals for my outcome variables. Essentially, these confidence intervals give me an estimate of my "effect sizes," which would be a range of plausible values for the population parameter. The 95% confidence interval for overall is 0.21–0.79, which suggests first-year students in the English class could improve their overall grade on papers between 0.21–0.79 points. The 95% confidence interval for macro is 0.15–0.77, which suggests first-year students in the English class could improve their overall grade on papers between 0.15–0.77. The 95% confidence interval for mid and micro is 0.13–0.77, which suggests first-year students in the English class could improve their overall grade on papers between 0.13–0.77.

But, these are only two *p*-values for the outcome variables overall and macro. So, I ran a permutation test of all my different variables for blank reasons. Why am I running a permutation test? The permutation test is a nonparametric way to test for differences between my samples (Collingridge, 2012; Fisher, 1935). One advantage to this approach is that I do not assume the type of distribution my data follows. If the data did not follow a normal distribution, and I only conducted a paired sample t-test, then my resulting *p*-value could be wrong from this initial assumption. A permutation test builds my distribution and then conducts the analysis, which is a more robust way to provide evidence against the null hypothesis. The second advantage of this approach is that I test all possible combinations of my control and treatment groups with my data. If I only test one combination of my observed data, it is more likely that the effects I see could be due to chance. By using a permutation test, I see all possible permutations of my control and treatment groups' test statistics, so I can be more confident that these observed effects are not due to chance.

There is one important assumption of permutation based tests. First, the observations in your data set are shufflable under the null (Kaplan, 2009). Thankfully, my data is permutable under the null hypothesis: why? I have subjects from a common pool—the students—and we think the treatment (professors' written comments) do not affect student writing. In the context of this study, this would mean teacher comments do not change the mean scores of overall and macro. If modifying the treatment doesn't change anything, then the numbers that identify which subject receives different treatments is arbitrary, so we can permute the values of the test statistic under the null. A consequence of this assumption is that the data in the permutation paired sample t-test have equal variance. Thankfully, my data has equal variance.

How does the permutation test work? It creates a sampling distribution by resampling different observations of the data (Gonick & Smith, 1993). More specifically, I can actually assign different outcome values to each observation while keeping the observed outcomes of my data. In the context of a permutation test, the null hypothesis states that the two distributions are equal. Put differently, the null hypothesis says treatment groups do not differ on the dependent varaible. By permuting all the different outcome values, I can see every combination of the possible alternative treatment conditions. Then, I can calculate all possible values of my test statistic by permuting all the different combinations of my control and treatment groups.

Stated otherwise, if my outcome is independent of the treatment, the permutation test allows me to see the difference in means for my observed data; afterwards, I can see these difference in means to compare all the possible differences in my data.

To clarify the permutation test, consider this example. Suppose, we have sample A and B with sample means a and b. The permutation test can inform us if the observed difference between the sample means is large enough to reject the null hypothesis—under the assumption that both groups have identical probability distributions. First, we calculate the observed differences between the two groups with our actual data and find the p-value. Second, we calculate the difference in sample means for every possible combination of group A and B. Then, we can see how likely it is that we reject or fail to reject the null hypothesis with every possible combination. By looking at all these different test statistics, we can calculate a final p-value. If the p-value is less than or equal to the selected α, we have strong evidence against the null hypothesis with our permutation test.

With my data, I have all of the students first paper grades for overall, macro, mid, and micro from 1 to 4. Each paper grade corresponds with a specific student, which represents a column of pre-treatment conditions in my experiment with s15students.Overall1 in Table 4. I also have all of the students second papers grades for overall, macro, mid, and micro, which are represented in the column s15students.Overall2 in Table 4. This column serves as the post-treatment condition. Then, I permuted all of the different values in s15students.Overall2 with the permutation test to obtain every possible combination of t-statistics between s15studentOverall1 and s15studentsOverall2 (Table 4). I repeated this same process for the macro outcome variable.

After running the two-tailed paired sample permutation test for overall and macro, I found a p-value of 0.004 for overall and 0.01 for macro, so I reject the null hypotheses (Table 5). These results suggest that students are improving the overall and macro quality of their papers, and these effects are most likely not due to chance. After running the paired sample t-test and permutation test, it is still possible that I am committing a type I error and the effects we see in student improvement are false positives. However, running both tests gives me more confidence that I am less likely to commit a type I error with my data. Moreover, I found a 95% confidence interval for overall as 0.13–1.28, which means first-year students in the English class could improve their overall grade on papers between 0.13–1.28 points. The 95% confidence interval for macro is 0.04–1.18, which suggests first-year students in the English class could improve their overall grade on papers between 0.04–1.18. These confidence intervals suggest that students could improve their starting grades on papers by one full letter grade or by nothing in the context of one semester. This means it would be highly unlikely for a student who get a “C” on the first paper in class would increase to an “A” level paper in these classes.

Table 4: Students' Paper Grades for Outcome Overall

Student	Overall Grade for First Papers	Overall Grade for Second Papers
1	2	2
2	1	2
3	2	2
4	2	2
5	3	4
6	3	3
7	1	3
8	2	3
9	3	2
10	3	3
11	2	3
12	2	2
13	1	2
14	3	4
15	2	3
16	2	2
17	1	2
18	1	1
19	2	3
20	2	3
21	2	3
22	3	2
23	2	2
24	2	3
25	3	3
26	2	3

N=26

A *p*-value only indicates evidence against our null hypothesis, but it doesn't tell us anything about our effect size. An effect size is the "difference between the average, or mean, outcomes in two different intervention groups" (Sullivan & Feinn, 2012, p. 279). We should calculate the effect size size because it describes the magnitude our treatment has on our outcome variable. If we only look at *p*-values, then we have evidence against the null hypothesis. However, we wouldn't know how meaningful of an effect our treatment has on our outcome variable. In the context of this paper, we wouldn't know how much student papers are improving from teachers' comments. So, what are the effect sizes for overall and macro?

In order to calculate the effect sizes of my outcomes variables, I decided to calculate Cohen's d. Before I explain why I chose Cohen's d, I need to define estimators and estimates. First, an estimator is "a rule that tells [us] how to calculate an estimate based on the measurements contained in a sample" (Weisstein, 2018b). An estimate is "an educated guess for an unknown quantity or outcome based on known information" (2018a). There are two types of estimates: interval estimates and point estimates. Intervals estimates take on a range of values, like confidence intervals, whereas point estimates take on a singe value like the standard deviation. Consider the following example to illuminate these distinctions. The American Veterinary Medical Association wants to publish a report on kennel dogs' weights in the US. The board members think this information could motivate people to adopt dogs if they spread the news in a nationwide campaign. However, they don't have the resources to measure the weight of every dog in the US. So, they decide to take a representative sample of kennel dogs' in all 50 states to find, on average, how much dogs weigh. In this context, the sample mean is an estimator for the population mean. So, we take the estimate of the estimator, the sample mean of the kennel dogs, and makes inferences about the population mean of all dog weights in the US that live in kennels. The two primary ways to discuss estimators are in terms of bias and varience. Bias is simply the difference between the "true" populationn parameter and the expected value of the estimator. Varience measures how far, on average, your estimates are from the *expected* estimate values.

Cohen's d is an estimator. In the context of my paper, I am taking two sample means—the grades of overall and macro for first-year students in my sample—and comparing them to the parameter: all undergraduate students who take a general education English course in their second semester. For independent t-tests, Cohen's d is found by calculating the mean difference between groups and dividing this difference by the pooled standard deviation (Cohen, 1988). Since Cohen's d calculates effect sizes by taking a standardized difference in means, I used this calculation to find the effect sizes between the students' first and second papers.

One reason I used Cohen's d was my sample size. Since my n $\geq$ 20, Cohen's d should produce a more accurate effect size than Hedges' g. As a result, the effect size for my calculation is less biased. Conversely, Hedges' produces less biased effect sizes when n $\leq$ 20 since it divides the difference in means by the pooled weighted standard deviations of the sample. Furthermore, Cohen's d assumes the variences of the populations are equal. In my study, the variances are equal since I am dealing with the same group of students. Thankfully, the variances of my population are equal, so this assumption is not violated. In the end, I used Cohen's d to calculate the effect sizes from the difference in means between the students' first and second papers for overall and macro.

I found an effect size of 0.71 for overall and 0.61 for macro (Table 5), but what do these effect sizes really mean? Conventionally, an effect size of 0.8 is considered large, 0.5 is medium, and 0.2 is small (Cohen, 1988). Generally, a d of 0.2 is an effect you can only observe with a carefully constructed study, whereas a d of 0.8 is an effect that someone can see with the naked eye. The meaning of an effect size, however, largely depends on context. For example, a d of 0.2 may be a large effect size if the context is preventing teenage suicide with nationwide prevention programs. Context and expertise is important when interpreting effect sizes: there is not a one-size fits all rule.

Given my small sample size, the effect sizes for overall and macro are fairly large. In practical terms, this means that students' writing abilities appear to be improving at a meaningful rate by the end of the semester. Many teachers anecdotally note improvement in their students' papers by the end of their courses, and these effect sizes support their claims. In the end, these effect sizes suggest that students are learning how to write better over the course of a semester in a meaningful way. Meaningful, in this context, would indicate that students write better thesis statements, organize their papers more efficiently, and use evidence in more compelling ways (see Appendix 3 for macro and overall rubric).

Then, I decided to find the error rate for my overall and macro level outcomes. I have included psuedo-code below to demonstrate how I find the error rates of my outcome variables overall and macro for all students.

```r
##Functions to Calculate False Positive Rates

#Create function to re-run experiment
newexperiment<-function(z,b){
  ## A function to randomly assign treatment within pair
  ## z = treatment assignment that is completely randomly
        ##assigned within block
  ## b = a block indicator
  unsplit(lapply(split(z,b),sample),b)
}

## doParallel will not work on all machines.
```

```r
library(foreach)
library(doParallel)
cl <- makeCluster(4) ## 4 cores on my machine
registerDoParallel(cl)

library(robustbase)
```

```
## Warning: package 'robustbase' was built under R version 3.4.2
```

```r
errratefn <- function(simulations,trt,outcome,block) {
  require(robustbase)
  outpaired <-  outcome - ave(outcome,block)
  output<-foreach(1:simulations,
                  .packages="robustbase",
                  .export=c("newexperiment"),
                  .combine='c') %dopar% {
    ## First make a new experiment with no
                    ##relationship between z and y
    ## since we know that there is no effect,
    ## then we can assess the false positive rate of
    ## lmrob
    newz <- newexperiment(z=trt,b=block)
    newzpaired <- newz - ave(newz,block)
    sim_p <- summary(lmrob(outpaired~newzpaired))$coefficients[2,4]
    return(sim_p)
  }
  return(output)
}

#False Postive Rate For Overall & Macro: Blocked For All Students
##Make block that includes all students.
  block <- c(rep.int(1, 26))

set.seed(12345)
results2a <- (errratefn(100,trt=s15students$Student,
              outcome=s15students$Overall2,block=block))
Overall_Error_Rate_Unblocked <- mean(results2a, na.rm = T)

results2b <- (errratefn(100,trt=s15students$Student,
              outcome=s15students$Macro2,block=block))
Macro_Error_Rate_Unblocked <- mean(results2a, na.rm = T)
```

I found error rates of 0.47 for overall and macro after all students are pooled together.

After finding the error rate, I calculated the power of overall and macro. Why did I calculate the power? The power is going to assess the probability of making a type II error; rephrased, my power shows the probability of my hypothesis test producing a false negative. The three factors that influence power are α, effect size, and sample size used to detect the effect size. Furthermore, power rates fall between values of zero and one; the higher the power, the less probable it is you make a type type II error. The lower the power, the more probable it is you make a type II error and falsely reject the null. For instance, a power of 0.8 means I have an 80% chance of correctly failing to reject the null (or a 20% chance of incorrectly failing to reject the null).

For my analysis, I ran 1000 different simulations of my data and found each of those power rates. I have included pseudo code below to show how I calculated my power.

```
#Function to calculate power.
  #simNum=Number of simultations
  #N=Sample Size
  #d=Effect Size (Cohen's d)
set.seed(12345)
t_func <- function(simNum, N, d) {
    x1 <- rnorm(N, 0, 1)
    x2 <- rnorm(N, d, 1)

    # run t-test on generated data
    t <- t.test(x1, x2, var.equal=TRUE)
    stat <- t$statistic
    p <- t$p.value

    return(c(t=stat, p=p, sig=(p < .05)))
        # return a named vector with the results we want to keep
}
```

Ultimately, I took the mean of all my power rates from these simulations to be more confident in my power rates. Instead of just running one simultation, running 1000 gives me more confidence that my power is more likely to be accurate.

I found a power rate of 0.72 for overall and 0.61 for macro after taking the average of all my simulations (Table 5). Overall, given my small sample sizes, these power rates aren't bad. Generally, studies should be designed to detect an effect 80% of the time and produce false negative roughly 20% of the time (Cohen, 1988.) We, of course, want higher power if possible, but with a sample size of 26 these powers are fairly good. This means that I most likely did not falsely reject the null hypothesis for these two variables.

What does all this analysis actually say about students learning how to write? It

Table 5: Pooled Classes Permutation Test, Effect Size, and Power

Level	95% CI	Permutation p-value	Cohen's d	Error Rate	Power
Overall	0.13 to 1.28	0.004	0.700	0.470	0.720
Macro	0.04 to 1.18	0.011	0.610	0.470	0.610

DF=25. Two-tailed test.

appears that students are learning how to write better from their classes. There is an average shift of 0.5 in the mean difference of students' overall paper quality, which means that most students started out at a "C" level of writing and ended the class with a final paper grade of "C+" or "B-." This is a huge shift in learning for one semester! Simply showing that students are learning how to write from professors comments is an important milestone. Moreover, most of this improvement is on global issues of writing related to thesis development, organization, logic, and use of evidence. Although one teacher emphasized grammar and citation style, both primarily commented on global features of the writing. Thankfully, students appear to be learning from their classes.

Seperate Classes Methodology & Results

However, I'm also interested in seeing if one style of commenting resulted in greater improvement of students writing ability. It's one thing to say that students are learning how to write from taking a class, but it's another thing to see if different styles improve student learning outcomes. So, I split the students into their respective classes and conducted four two-tailed paired sample t-tests to see if different classes showed different levels of improvement with α=0.05.[2]

After subsetting the two classes, my sample size shrunk. My class 1 n=16 and class 2 n=10. However, despite these small sample sizes, I still reject the null hypothesis for class 1's overall and macro outcomes (Table 6). Conversely, I fail to reject the null hypothesis for class 2's overall and macro outcomes with p=0.052, which is just on the threshold of statistical significance.

For class 1, I find point estimates of 0.5 for overall and 0.44 for macro. These point estimates suggest that students improved by roughly half a letter by the end of the course in terms of their overall paper and global features of the writing. The 95% confidence interval for class 1 on overall is 0.11–0.89 and 0–0.87 for macro. This shows the range of potential values that students could improve their grade by in class 1. In extreme cases, the students could improve by almost a full letter grade or

[2] All the assumptions of the paired-sample t-test are satisfied in this context.

Table 6: Separate Classes Paired Sample T-Test

Class and Level	95% CI	t-statistic	p-value	Mean Difference
Class 1 Overall	0.11 to 0.89	2.740	0.015	0.500
Class 2 Overall	-0.01 to 1.01	2.240	0.052	0.500
Class 1 Macro	0 to 0.87	2.150	0.048	0.440
Class 2 Macro	-0.01 to 1.01	2.240	0.052	0.500

Class 1 DF=15. Class 2 DF=9. Two-tailed test for both classes.

show virtually no improvement in their writing ability by the last paper (see Table 6). Unfortunately, I failed to reject the null hypotheses for class 2 overall and macro, so I won't discuss their point estimates or confidence intervals in great detail (see Table 6).

I decided to run a two-sided permutation test for these two different classes and outcome variables. Since the permutation test does not assume the normal distribution, I wanted to construct my distribution from the data and see how robust my previous results are from the paired sample t-test. [3] After I ran my permutation test, it appears that I fail to reject the null hypothesis in most cases (Table 7), but students in class 1 appear to be improving their overall quality of papers with p=0.035. The 95% confidence interval for Class 1 Overall represents the range that students could improve from the class. Failing to reject the null hypotheses is most likely due to my smaller sample sizes and my study being underpowered. In order to find the power of my studies, I need to find the effect sizes of the different outcome variables.

I used Hedges' g to find the effect size for the two different classes and their respective outcome variables (Table 4). One reason I chose Hedges' g as my point estimate are my sample sizes. When n $\leq$ 20, Hedges' g produces less biased effect sizes since it divides the difference in means by the pooled and weighted standard deviations of the sample. Dividing by the pooled and weighted standard deviation helps account for the small sample size by getting a more precise measure of the standard deviation of the small sample and removes a small positive bias (Hedges, 1981). If I were to use Cohen's d for these different calculations, the effect sizes could be more positively biased given the small sample sizes.

For class 1, I found error rates of 0.48 for overall and macro. For class 2, I found error rates of 0.49 for overall and 0.47 for macro. After I found my error rates, I decided to find the power rate for the two classes and their outcome variables. I ran 1000 different simulations of my data and found 1000 different power rates like I did previously. Afterwards, I took the mean of all my power rates, which gave me more confidence they were accurate.

[3] All the assumptions of the permutation paired-sample t-test are satisfied in this context.

Unsurprisingly, I have low power rates of 0.450 for Class 1 Overall as well as 0.3 for Class 2 Overall, Class 1 Macro, and Class 2 Macro (Table 4). This is unfortunate. It's highly possible that some type of effect is occurring with these students, but the test power is small, so I'm not detecting any noticeable effects. I do, however, detect an effect for improvement in the first class' overall paper quality. Hedges' g for Class 1 Overall is 0.67 with a mean difference of 0.5. In practical terms, this means a student's first paper would be graded as a "C", and their final paper grade would be a "C+" or "B-." Moreover, given the effect size of 0.67, it is entirely possible that teachers could see their students improve their writing ability over the course of the semester.

Table 7: Separate Classes Permutation Test, Effect Size, and Power

Class and Level	95% CI	Permutation p-value	Hedges' g	Error Rate	Power
Class 1 Overall	-0.07 to 1.41	0.035	0.670	0.480	0.450
Class 2 Overall	-0.29 to 1.64	0.125	0.680	0.490	0.300
Class 1 Macro	-0.21 to 1.26	0.092	0.520	0.480	0.300
Class 2 Macro	-0.29 to 1.64	0.125	0.680	0.470	0.300

Class 1 DF=15. Class 2 DF=9. Two-tailed test for both classes.

Part III

Discussion

When all students from both classes are pooled together, they show improvement across outcomes overall and macro with fairly substantial effect sizes. The statistical analysis for all students would suggest that students, in fact, learn and improve their writing abilities by 0.5 on overall and 0.46 for macro. Essentially, all students, regardless of the commenting style, improve the overall quality of their paper. Moreover, they also improve macro features of their papers like organization, evidence use, and strength of argument. In grading terms, most students in the course started at a "C" level and turned in a final paper at a "C+" or "B-" level. Given the difficulty of the final assignments—writing a research based paper with a thesis statement, clear supporting evidence, cogent organization, concise topic sentences, and other factors—this is a substantial learning outcome. (This finding is heartening because students are learning! It would be disheartening if they didn't learn.)

After I separate the students into separate groups and conducted my analysis, the only variable that was statistically significant with the paired sample t-test and permutation test was Overall Class 1. All other variables were not statistically significant with the paired sample t-test and permutation test. These results could suggest that professor 1 gave comments that demonstrate greater improvement in their students' overall papers, whereas professor 2 gave comments that result in less improvement for their students' overall papers. Put differently, I was able to reject the null hypothesis with a small sample size for Class 1 Overall, but I failed to reject the null hypothesis for Class 2 Overall with a small sample size. This finding could suggest professor 1 gave comments that helped students learn more, whereas professor 2 could have given comments that resulted in less student learning. However, given my low power rates, it is highly possible there is an effect going on with Class 2 Overall: I could be commiting a type II error.

If the effects for Overall Class 1 are not due to the slightly larger sample size in comparison to Overall Class 2, this would suggest professor one's method of commenting with more evaluative praise improves student writing at a faster rate than professor two's method of commenting. Why would this be the case? Since most of the students are in their first-year, I argue they have a difficult time understanding jargon and incorporating these types of comments into their own writing. Professor one mostly gives evaluative comments that simplify if the student wrote a great topic sentence or should try improving their thesis. This explicit evaluation results in greater improvement since the students have an easier time understanding the comments, and are more capable of implementing this feedback into their papers. Conversly, professor two writes with more technical terms, which could explain why this group of students showed less improvement. At the end of the day, students appear to understand evaluative comments better than technical terminology.

Ultimately, the practice of teaching writing to students is lore based. I define lore as group knowledge or traditions that are passed on person-to-person. In the context of teaching writing to students, lore boils down to a simple idea—this is what I do when I teach writing to students. Driscoll & Perdue (2012) offer a meaningful quote about lore in the context of writing centers: "While it [lore] is often marketed as research and inhabits a substantial place in *WCJ* [the Writing Center Journal], this kind of scholarship offers little more than anecdotal evidence, one person's experience, to support its claims" (p. 16). One of the top journals in writing center praxis relies on anecdotal claims to support teaching strategies in writing centers. Consequently, the ability of *WCJ* to conduct reproducible research is highly limited. Driscoll & Perdue (2012) further elaborate, "While *WCJ* has published many articles that discuss data, we would classify only sixteen percent of its articles as replicable, aggregable, and data-supported research. . . .Clearly, the story of RAD research in *WCJ* is one of yet-to-be-met potential" (p. 29). Although these quotes focus on *WCJ*, the main idea translates to how professors teach general education English courses. Professors adopt specific lores and values that their teachers taught them when they learned how to write; often, these inhereted ways of teaching writing are left unchecked. As a result, it is possible that certain types of students may benefit more from data-drive decision making in their learning rather than a lore based approach.

Future Directions

Future directions for this research could include text analysis or machine learning algorithims that analyze a larger number of professors written comments. As long as professors write comments in a word document or PDF, this could be an effective way to calculate larger n-sizes and cut down the time it takes researchers to code different dimensions of the professors' comments.

Furthermore, students enter college from vastly different backgrounds. Future studies could look at students from traditionally marginalized backgrounds and see if certain commenting styles help them learn better. For example, students that are first generation are usually less prepared for the rigors of college, and they may benefit more from a different commenting style than other students.

Moreover, this study did not include non-native English speakers. It is entirely possible that different commenting styles would be more effective for non-native English speakers. Different countries have different styles of writing, and these different conventions of writing could influence what commenting styles are most effective with this group of students.

Ultimately, the primary strength of this study is its incorporation of statistical methods to pedagogical writing—a highly lore based approach. Most research in writing studies focus on theories of teaching writing, and this approach calls for more data-driven decision making to prioritize student learning.

References

AERD Statistics. (2018). Dependent t-test using SPSS statistics. Retrieved from https://statistics.laerd.com/spss-tutorials/dependent-t-test-using-spss-statistics.php.

Beason, L. (1993). Feedback and revision in writing across the curriculum classes. *Research in the Teaching of English*, *27*(4), 395-422.

Boneau, C., A (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin, 57*(1): 49–64.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Collingridge, D. S. (2012). A Primer on quantitized data analysis and permutation testing. *Journal Of Mixed Methods Research*, *7*(1), 81-97.

Driscoll, D., and Perdue S. (2012). Theory, lore, and an analysis of RAD research in the writing center journal, 1980– 2009. *The Writing Center Journal*, *32*(1), 11–39.

Fisher, R. (1935). The design of experiments. 1935. Oliver and Boyd, Edinburgh.

Hsu, H. and Lachenbruch, P. A. (2014). Paired t Test. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels). doi:10.1002/9781118445112.stat05929

Gonick, L. and Smith, W. (1993). The cartoon guide to statistics. HarperPerennial New York, NY.

Kaplan, D. (2009). Statistical Modeling: A Fresh Approach.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107-128.

Koc, W. E., Koncz, J. A., Tsang, C. K., & Longenberger, A. (2016). NACE job outlook. National Association Of Colleges And Employers. Retrieved from https://www.mccormick.northwestern.edu/career-development/documents/getting-started/job-search/NACE%20Job%20Outlook%202016.pdf

Murphy, C, & Sherwood, S. (1995). The st. martin's sourcebook for writing tutors. New York: St. Martin's Press.

NCSS. (2018). Paired t-test. Retrieved from. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Paired_T-Test.pdf

Sommers, H. (2013). Responding to student writers. Boston, NY: Bedford/St. Martin's.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. http://doi.org/10.4300/JGME-D-12-00156.1l

Weisstein, Eric W. (2018a). “Estimate.” From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/Estimate.html

Weisstein, Eric W. (2018b). “Estimator.” From MathWorld—A Wolfram Web Resource. http://mathworld.wolfram.com/Estimator.html

Appendix 1

Coding & Identification System for Professors Comments on Student Writing

Part 1: This section is attempting to identify the professor's purpose in making the comment. The individual looking at the paper should identify the best fit from the items listed below.

1. Problem Detecting: Indicates a problem, concern, or error (e.g., "SP" "What?" "Coherence" "This isn't quite accurate.").
2. Advising: Gives general options or direction but does not offer the actual deletion, punctuation, or language needed (e.g., "Consider deleting some of this." "This would be convincing if you addressed the opposition." "Can you explain more clearly?"). Might explain why change is needed.
3. Editing: Indicates a problem and supplies the actual deletion, punctuation, or language needed (e.g., "Put comma here." "Drop this." "Would 'person' be a better word?"). Might explain why change is needed. Praising: Shows approval (e.g., "Great!" "You've hit upon some-thing.").
4. Praising: Indicates approval, which may or may not include a suggestion. (e.g., "Good! Add more of this!" "Nicely said but content needs work.").
5. Describing: Describes text or paraphrases in an apparently neutral way (e.g., "The paper supports ethics in journalism").
6. Topical Commenting: Reflects on the subject (rather than on the writing) in an apparently neutral way (e.g., "This makes me think about values." "There is a law dealing with this problem you raise." "This is a popular topic").
7. Other: Does not fit any of the above. Please specify the professor's purpose in making the comment by offering a different reasoning not offered above.

Part 2: This section is attempting to identify the style of commenting that the professor is utilizing. The individual looking at the paper should identify the most accurate style of commenting from the items listed below. Comments can be considered in one of three categories: directive, nondirective, and corrective.

1. Directive. Offers feedback that advises the students in order to correct a problem or directly identify a positive component of their writing, but does not directly make the change. One example a professor may write would be, "In future papers you should think about developing your thesis a little more." Another example an instructor may write is, "Excellent word choice and thesis!"

2. Nondirective. Offers feedback that poses questions, indicates neutral thoughts about the paper, or makes marks that indicate a good idea or problem with the paper on some level, but in such a way that allows the writer of the paper to figure it out. (How do you feel about this topic? What do you enjoy about your paper?)
3. Corrective. Corrects a problem explicitly with crossing out, adding punctuation, or other such methods that indicate to the writer a problem exists and provides the solution. Something a professor might do to indicate this would be crossing out a semicolon and indicating that it should be a comma instead.
4. Evaluative. Offers explicit judgment or identification of the student's or paper's strengths and weaknesses. Gives praise or criticism of the paper in general. Writes comments that evaluate the student or work. Writes that the paper or a specific portion of the work is "good" or "bad." For example, a professor may write, "This paper is terrible and needs work." This would be considered evaluative.

Part 3: This section is attempting to identify if the professor is offering comments on the macro, mid, or micro-level. The individual looking at the paper should attempt to identify one level of writing the comment is directed at and the specific article the comment is attempting to identify. For example, if the researcher decides the professor's comment is macro-level oriented, the research would follow up with specifying that it was a thesis level problem.

Macro-Level Features (Macro=3):

Article 1. Thesis A paper articulates a thesis with a clear and specific topic, an arguable claim, and conveys the implications of his or her work in a coherent manner.

2. Organization Paper builds in a manner that makes logical sense.

3. Evidence Evidence used throughout the paper is effective and suitable to the argument(s) presented.

4. Depth and Thoroughness of Discussion Paper is thorough.

5. Logic and Reasoning The paper is logical and builds upon sound reasoning.

Mid-Level Features (Mid=2):

Articles 6. Syntax and Overall Composition of Sentences Sentences vary their construction, are not repetitive, and communicate meaning effectively.

7. Effective Transitions Transitions are used effectively throughout sentences.
8. Topic Sentences Sentences effectively convey the paragraphs argument or area of interest.
9. Sentence Conciseness Sentences are concise and non-verbose.

Micro-Level Features (Micro=1):

Articles 10. Grammar Follows conventions of grammar and usage that is appropriate in an American academic setting.

11. Punctuation Punctuation is used in a manner that conveys intended meanings.
12. Spelling Words are spelled correctly and appropriately.
13. Word Choice Diction and tone reflects an appropriate meaning of each word.
14. Capitalization Words are appropriately capitalized.
15. Citation/Works Cited Page Citation style follows conventions that are appropriate to the specific style of citation on a micro-level scale. More concerned with mechanical features of citation and works cited page.
16. Other Please specify what other thing you are talking about.

Appendix 2

Table 8: Frequency of Purpose Comments

Professor	Purpose	Frequency
1	Problem Detecting	48
2	Problem Detecting	49
1	Advising	46
2	Advising	92
1	Editing	14
2	Editing	16
1	Praising	86
2	Praising	19
1	Describing	7
2	Describing	0
1	Topical Commenting	0
2	Topical Commenting	2

Class 1 N=201. Class 2 N=178.

Table 9: Frequency of Purpose Comments

Professor	Style	Frequency
1	Directive	75
2	Directive	117
1	Nondirective	28
2	Nondirective	26
1	Corrective	14
2	Corrective	15
1	Evaluative	84
2	Evaluative	20

Class 1 N=201. Class 2 N=178.

Appendix 3

Table 10: Frequency of Level Comments

Professor	Level	Frequency
1	Micro	30
2	Micro	65
1	Mid	43
2	Mid	26
1	Macro	128
2	Macro	87

Class 1 N=201. Class 2 N=178.

(4: Excellent. 3: Great. 2: Good. 1: Poor)

Part 4: This section is attempting to assign a score from 1-4 for the overall macro-level composition of the paper.

4 Macro-Level

Demonstrates a sophisticated awareness of the audience and their expectations. Clearly indicates purpose for communicating, achieves that purpose, and contributes to greater understanding of the topic. Persuasively supports observations or claims with skillfully chosen examples, with compelling evidence from authoritative sources, and with exhaustive reasoning. Presents ideas in a outstandingly controlled, fluid, and logical manner, both across the work and within individual paragraphs or sections.

3 Macro-Level

Demonstrates a consistent awareness of the audience and their expectations. Clearly indicates purpose for communicating and achieves that purpose. Supports observations or claims with sufficient and appropriate examples, with evidence from authoritative sources, and with thorough reasoning. Presents ideas in a logical and easy to follow manner, both across the entire work and within individual paragraphs or sections.

2 Macro-Level

Demonstrates inconsistent awareness of the audience and their expectations. Shows an incomplete sense of purpose or needs further work to achieve purpose successfully.

Table 11: Frequency of Article Comments

Professor	Article	Frequency
1	Thesis	16
2	Thesis	12
1	Organization	38
2	Organization	18
1	Evidence	34
2	Evidence	19
1	Thoroughness of Discussion	3
2	Thoroughness of Discussion	19
1	Logic and Reasoning	39
2	Logic and Reasoning	19
1	Syntax	19
2	Syntax	12
1	Transitions	8
2	Transitions	9
1	Topic Sentences	1
2	Topic Sentences	1
1	Sentence Conciseness	14
2	Sentence Conciseness	3
1	Grammar	15
2	Grammar	10
1	Punctuation	5
2	Punctuation	3
1	Spelling	4
2	Spelling	9
1	Word Choice	4
2	Word Choice	7
1	Capitalization	0
2	Capitalization	3
1	Citation/Works Cited Page	1
2	Citation/Works Cited Page	34

Class 1 N=201. Class 2 N=178.

Supports observations or claims with insufficient or inappropriate examples, with evidence that is inadequate or does not come from authoritative sources, or with underdeveloped reasoning. Does not always present ideas in a logical and easy to follow manner. Paragraphs, sections, or sentences may appear out of sequence or not relate to those immediately before and after.

1 Macro-Level

Demonstrates little awareness of the audience and their expectations. Does not achieve purpose for communicating. Does not support observations or claims or provides only minimal support. Evidence is inadequate or does not come from authoritative sources and reasoning is faulty or insufficient. Does not present ideas in a logical or easy to follow manner.

Part 5: This section is attempting to assign a score from 1-4 for the overall mid-level composition of the paper.

4 Mid-Level

States ideas in clear, concise sentences and demonstrates a masterful and distinct control of sentence structure. Uses words skillfully and at a consistent level of diction suited to the purpose.

3 Mid-Level

States ideas in clear, concise, and varied sentences. Uses words appropriately and at a consistent level of diction suited to the purpose.

2 Mid-Level

Demonstrates recurring problems with clarity, wordiness, or lack of variety in sentence structure. Uses some words inappropriately or at inconsistent levels of diction.

1 Mid-Level

Demonstrates significant problems with clarity, wordiness, or lack of variety in sentence structure. Frequently uses words inappropriately or at inconsistent levels of diction.

Part 6: This section is attempting to assign a score from 1-4 for the overall micro-level composition of the paper.

4 Micro-Level

Demonstrates comprehensive knowledge of appropriate communication standards. Contains few if any errors in grammar, punctuation, spelling, or capitalization. For oral communication, is consistently delivered clearly and fluently.

3 Micro-Level

Demonstrates significant knowledge of appropriate communication standards. May contain occasional errors in grammar, punctuation, spelling, or capitalization. For oral communication, is generally delivered clearly and fluently.

2 Micro-Level

Demonstrates inconsistent knowledge of appropriate communication standards. Contains some recurring errors in grammar, punctuation, spelling, or capitalization. For oral communication, shows some problems with clarity and fluency.

1 Micro-Level

Demonstrates little knowledge of appropriate communication standards. Contains significant and recurring errors in grammar, punctuation, spelling, or capitalization. For oral communication, shows frequent problems with clarity and fluency.

Part 7: This is the final section of the paper that attempts to judge the overall quality of the paper as a complete product. Utilize the rubric listed below when helping you to make your determinations.

4 Paper

A 4 paper is significant in purpose and sensitive to the audience for which it is written. It presents a logical thesis, coherent structure, and paragraphs organized by a controlling idea. Transitions are effective both between and within paragraphs. Each paragraph is fully developed through analysis and examples appropriate to the thesis. Not only is the paper well supported, but the style is also clear with little awkwardness or ambiguity. Sentences show variation, and diction is sensitive and precise. Furthermore, few, if any, mechanical errors exist. The total effect is that of a fresh, personal, and provocative paper.

3 Paper

The 3 paper contains some but not all the attributes of the 4 paper. This paper is also significant and well argued, but some of the arguments are not as well supported as they could be. Even though it might have some minor lapses in its reasoning, the paper still contains a worthy thesis, logical organization, and developed paragraphs with effective transitions. Generally, the sentence structure will be varied and correct, but a few mechanical errors or awkward sentences might exist. Diction will be correct but not as sensitive or sophisticated as in the 4 paper.

2 Paper

The 2 paper is also organized around a thesis statement, but the thesis may not be as clearly defined as in higher quality papers. The topic expressed in the thesis may be trivial or contain assumptions which the writer never recognizes or renders palatable to the reader. The organization is obvious but possibly formulaic, and transitions may not be smooth. Although the paper focuses on a topic, it may have problems with paragraph unity, development, or adequate support. Most of the sentences are correct, but some might be choppy, repetitive, or lacking in variety.

1 Paper

The 1 paper has no central idea or one that is too general to be developed. Paragraphs are not logically connected, and transitions do not generally exist. Development is inadequate with poor balance between general ideas and specific support. Sentences are often ungrammatical and word choice faulty. As a result, the paper fails to present a conscientious inquiry into a problem and seems to disregard its audience.