An Examination of Electronic and Traditional Instructor Feedback:
Quantitative Comparison of the Discourse of Marginal Comments

by
Way Jeng

Submitted in Partial Fulfillment of the Requirements
for the Degree of
Master of Arts
in the
English
Program

YOUNGSTOWN STATE UNIVERSITY
May, 2010
An Examination of Electronic and Traditional Instructor Feedback:  
A Quantitative Comparison of the Discourse of Marginal Comments

Way Jeng

I hereby release this thesis to the public. I understand that this thesis will be made available from the OhioLINK ETD Center and the Maag Library Circulation Desk for public access. I also authorize the University or other individuals to make copies of this thesis as needed for scholarly research.

Signature:

Way Jeng, Student

Approvals:

Kevin Ball, Thesis Advisor

Jay Gordon, Committee Member

Jeff Buchanan, Committee Member

Peter J. Kasvinsky, Dean, School of Graduate Studies & Research
ABSTRACT

This study undertakes a quantitative, comparative study of instructor feedback on essays. The ubiquity of word processors in the classroom for both students and instructors has led to new situations where instructors give students feedback through comments directly embedded in electronic files, generated through Microsoft Word.

247 essays were collected in total, across two distinct phases of the study. 93 essays were collected in Phase One, with the remaining 154 essays collected in Phase Two. These essays reflected comments written by a total of 14 instructors, of whom four wrote comments electronically, six wrote comments traditionally (via hand writing on paper copies of essays), and four wrote comments in both styles. These comments were coded according to a rubric adapted from Kwangsu Cho et al. (2006) to determine broad categories such as Directive, Non-Directive, Praise, Criticism, and Off-Topic. Measures such as average comment length and number of comments per page were also examined.

Early findings in Phase One indicated that electronic comments tended to generate more comments per page, longer comments by word count (approximately 50% in both categories when compared to traditional comments), more Directive comment, and fewer Non-Directive comments. However, Phase Two research showed that only the comment length proved consistent across all instructors.

Re-framing the study in terms of individual instructor profiles offers a way to utilize these research techniques in a way that provides a detailed description of how each instructor comments in each style, thereby giving those instructors the ability to tailor their comments to each specific student.
Introduction

It turns out that grocery stores sell whipped cake frosting these days. I discovered this fact quite by accident when I went to a grocery store for confectioner's sugar. Rather than buy sugar to make into frosting and then whip with a mixer, I could buy a ready-made can of the finished product. It's a brave new world, where technology has made baking faster and easier than ever before. Apparently I'd been doing things the old-fashioned way without even knowing it. There was no way to know how long I'd been behind the times because the world had changed while I wasn't looking. It's conceivable that if I'd never walked down the icing aisle, I might have never known that the option was available to me.

This problem isn't unique to baking. Technology has entered into composition classrooms around the nation, and it's difficult to imagine them going away. The effect of computers on the classroom for students over the last ten years has been fairly pervasive. Typed papers are the norm, and the venerable typewriter has gone the way of obsolete technology. Indeed, Catherine Davidson notes in her study of integrating computers into an ESL class that one of the first uses of computers in her classroom was simply as a typewriter (92). This was also observed in Ilana Snyder's introduction of computers into a writing classroom (176).

Yet the uses of computers in the classroom go far beyond simple typing uses. Beyond providing students with new ways to research topics and aggregate data, the act of computer mediation of writing and sharing work has more far-reaching effects on students' work. Kate Kiefer and Mike Palmquist found that teachers who moved between
traditional classrooms and computer-based classrooms noticed that the focus on the computers, rather than the instructors, made students more comfortable. Students spent more time composing during class, and had an easier time revising their work ("Adapting to the Classroom Setting" 145-148; "How Does Access to a Computer Network Affect Writing Students' Interactions With Peers and Teachers?" 363).

Yet for all the effort that universities have put into making computers available in classrooms and campuses (installing computers and projectors in classrooms, making computer labs available for students without laptop computers, and setting up wireless networks), relatively little emphasis has been placed on integrating computers into instructors' feedback on student essays.

Modern word-processing programs such as Microsoft Word now have the ability to embed marginal comments inside of electronic files (Thompson 391). These so-called electronic comments (comments generated electronically through software such as Microsoft Word, but possibly including other software such as OpenOffice or Adobe Acrobat) have the potential to replace so-called traditional comments (comments written by hand, usually onto paper copies of essays).

It's not clear why instructors haven't widely adopted computers as part of their annotation practices. Stephen Carmichael and Peg Alden note that while most language disability students are quick to embrace computers, instructors continue to comment on student papers the same way they have for centuries: by writing marks on paper copies. Indeed, many instructors require or accept electronic submission of essays or other work, yet print out physical copies for evaluation (43-44).
Certainly there has been no lack of thinking about feedback over the last thirty years. Grading and commenting on student texts is a long, exhausting process that takes up large amounts of instructors' time and attention (Sprinkle 273). Much of the literature has been devoted to analyzing and considering the commentary feedback that instructors provide, as well as the related effectiveness for students. Sharon Mitchler's compilation of several studies indicates that students prefer close readings of essays and specific comments for how to make papers easier to understand, and additionally preferred comments before as well as after due dates (447-448). Unfortunately, research has also spoken out against commentary. Anthony Edgington notes that marginal comments, comments written or typed in the margins of papers, are seen as fragmented and unhelpful by students (290-291).

There has been very little written explicitly in defense of traditional comments. Elizabeth Hodges defends hand-written comments in her article 1997 article “Negotiating the Margins: Some Principles for Responding to Our Students’ Writing, Some Strategies for Helping Students Read Our Comments.” She suggests that instructors “write well” in the margins for better student understanding (81). Hodges defines “well” as making “clear, connected, useful, respectful comments” (81). Hodges does not speak in terms of instructors writing in legible handwriting but more in terms of the content of the commenting. Stephen Carmichael and Peg Alden do mention several other reasons instructors choose to use traditional commenting. These include creating a more personal connection to the student and being able to make the essay more portable without the need of a computer to make the comments (53). Anecdotally, it seems that some instructors are reluctant to electronically-generate feedback because doing so requires an
instructor to read a computer screen for several hours at a time to grade essays, and given the large amounts of time necessitated (we can imagine an instructor grading essays for three or four sections of Writing I!), this is an understandable sentiment. Instructors who experience eyestrain after a few hours will not want to use a feedback system that will require dozens of hours in front of a computer.

While these reasons may resonate with many instructors, the most likely reason why instructors have not switched over to electronic commenting is institutional inertia. Quite simply, the system is not obviously broken, so there's no need to change it. Davidson noted that when computers were introduced to her classroom, she feared that they would be "another obstacle to engaging student" (89), and Snyder noted that introducing teachers to computers in a writing class resulted in minimal change to the course (176).

Much has been written in favor of a switch to electronic commenting. Stephen Carmichael and Peg Alden noted that there are many benefits available to instructors who choose to provide electronic comments, such as expressing ideas and suggestions completely and taking less time to write comments. Instructors also have the ability to color-code comments so that students can easily recognize categories of comments, or might even add hyperlinks to essay comments that would direct students to more comprehensive resources (45-51). The portrayal of computer-based instruction as somehow qualitatively changing the possibilities of how teachers can view feedback and instruction have been around for quite some time now. As early as 1995, Signe Marie Sanne suggested that computer programs designed to teach students grammar would be
superior to textbooks because hyperlinks could provide students with detailed
explanations only a single click away (293).

Students may benefit from increased legibility, which will make the revision
process more efficient (Carmichael and Alden 46-47; Byrne 103). It's possible to edit
documents faster on a computer, especially with the ability to cut and paste (Byrne 103-
104). Philip Denton et. al. observed that students report equal or slightly higher
satisfaction with electronic comments, compared to traditional comments (495-498), and
Ros Byrne found that 75% of his students reported strong agreement, and 25%
agreement, with the statement "the teaching staff gave me constructive feedback on my
assessed work" on a course evaluation when the course was taught using electronic
feedback, though there was no traditionally-commented class to serve as a control (Byrne
104).

As such, there is compelling reason to believe that the idea of electronic
commenting has been around for some time, though it appears as though electronic
commenting is still rare in composition classrooms. Andrea Lunsford and Karen
Lunsford found in their survey of 877 student papers from numerous institutions showed
that only 56 had comments generated through some electronic means (794). This finding
generally agrees with the relative scarcity of research that examines electronic
commenting specifically, rather than instructor feedback generally. The reasons why
electronic commenting is still relatively rare is doubtless a complicated question. Byrne
noted in 1997 that limited computer compatibility and availability might be
disadvantages for electronic commenting (105), yet today we would be hard-pressed to
find faculty offices without computer, a campus without multiple computer labs, or even
adjunct faculty who possess neither a laptop to bring to work nor a computer at home (and indeed, many have both). So, too, has the de facto spread of Microsoft Word, and its spread to the Windows and OS X (for Apple computers) operating systems rendered compatibility a moot point for the vast majority of instructors. Another possibility is that instructors simply are not aware of the possibility, though several how-to articles have appeared in the literature, making such options available to instructors (Thompson 391; Holmes).

But even if electronic methods of commentary are not the norm now, it is important that we begin examining the differences between paper and electronic comments to determine if one form favors the creation of comments that instructors or students prefer. This question transcends the question of if instructors are providing electronic comments and instead asks if instructors should provide electronic feedback.

Thus, we are left with a burning question: Is there in fact a difference between traditionally-written comments and electronically-written comments? If the actions of students are changed by virtue of interacting with a computer, and mediating experience through that instrument, then it's entirely possible that instructors would find the similar (though not necessarily equal) effects. Amazingly, none of the research I was able to uncover answered this question. There was plenty of theory and thought surrounding the issue, but very little dedicated to actually measuring the real-world observations of electronic comments compared to traditional comments. Depending on the answers, the long-term implications on student revision could be extremely important. For example, understanding the ways that instructors naturally comment might suggest ways of
selectively using those comments to generate certain kinds of comments that students
would find more helpful while revising.

The questions of what the differences in commenting styles were, and which one
was better for general use in the composition classroom served as the foundation of this
research. My search for answers lasted about a year and a half, and the three chapters that
follow chronicle that journey. Chapter One begins with the initial research as I imagined
it during my Methods of Composition Research course my second semester in graduate
school. During that time I generated the core of the study's methodology, conducted a
modest data collection, and analyzed the data for results. Chapter Two describes the work
I did after the class in an effort to expand the data collection as well as additions that I
used to refine the methodology, as well as the frustrations that I experienced and the
problems that became evident in the study as it expanded.

Ultimately, though, those same research questions would give me much cause for
concern. In the third chapter, I examine the problems that occurred in Chapter Two and
reflect on the study to draw some conclusions about the research as a whole. In this
section I reflect on what it really meant for one commenting style to be "better" than
another, and I examine the ways that my inquiry had been shaped by unconscious
ideologies of research that had influenced the original research questions I created as well
as the methodologies that I used.
Chapter One: Early Beginnings:
Generating Initial Research Questions and a Methodology

The first thing that I wanted to do was make absolutely certain that I had numbers to work with. I’d heard some anecdotal accounts of instructors choosing one or the other style while taking some coursework in composition, but there was something reassuring about having raw numbers to examine. Instructors can sit in their offices and talk about how "connected" they feel to the essays, or discuss how their comments "come more naturally" when hand-writing or go faster electronically (perhaps with an accompanying claim of speed or additional comments), but I wasn't particularly moved by any of these accounts. Having these thoughts is very well and good, but I just couldn't bring myself to trust "eyeballed" self-reports.

My dissatisfaction went beyond simple self-reporting bias concerns. If there were differences, I wanted to know what they were. If one of the methods tended to generate more comments for the students, then it was important to know how much more was generated. If the difference was fairly minimal, say within 5%, then any advantage to be gained by switching commenting styles might not be enough to overcome instructor preferences. Added to that, I suspected that individual instructors would have widely differing commenting styles. Some would naturally lean towards longer comments, while others might generate many more but shorter comments. Judging such widely varying data would be very difficult without calculating statistics to find trends inside of the "noise" generated by a pool of very diverse data. Once those trends had been identified, I (or other researchers) would be able to go back and start to examine the trends in further
detail. The first task was to form a baseline of activity that future research could use to examine the various possibilities for instructor commenting.

Once I'd settled on a quantified study, several more factors fell into place naturally. Because the study aimed to find the real-world comments written on student essays, it made the most sense to collect actual commented essays from composition classes. To gather these essays, I approached several graduate assistants at Youngstown State University. Drawing from this limited population allowed me to control for any experience bias. The main idea I wanted to get at was whether or not the method of commenting itself had an effect on the generation of comments. Drawing data from the adjuncts or professors in the department might have meant comparing instructors with a few years of experience against others who might have been teaching for fifteen years or more. There was also the possibility that some adjuncts might have an MA, though others might have held an MFA, and the full-time faculty generally held PhDs. Eliminating these experience and educational biases was very easy by limiting data to a smaller population.

Ultimately, seven graduate assistants and one adjunct from the Youngstown State University English Department volunteered to take part in the study. The graduate assistants were in either their first or second semester of teaching, and none of them possessed an advanced degree of any kind. The one adjunct who participated was a newly-graduated graduate assistant working as an adjunct for the first time. Though he possessed an MA, his education and experience was generally equivalent to the other participants.
This pool of subjects proved very conducive to eliminating potential biases. All of the participants were between the ages of thirty and twenty-one. None possessed a teaching license, although two had worked in writing centers as undergraduates. On the whole it was a fairly homogenous group. Having a group who was attending the same school, going through the same composition teacher training, allowed me a fantastic chance to look at the effects of the commenting mode itself, unfettered by other concerns.

Getting the necessary permissions went relatively smoothly. All of the instructors who participated in the study were assigned non-identifying (though gender-appropriate) pseudonyms. All were informed of the study's purpose and made aware of their right to withdraw from the study. Each instructor signed notices of informed consent. All participating instructors then approached the section(s) of Writing I that they taught and distributed notices of informed consent to their students to collect the essays for research purposes. All students who donated essays to the study were informed of the study's purposes, and signed statements of Informed Consent. Names of these students were removed from the collected essays to ensure the confidentiality of their identities. The Youngstown State University Human Subjects Committee reviewed the study's protocols to ensure ethical treatment of subjects.

Each instructor provided me with one essay from each student who volunteered to take part in the study. Due to each instructor's preference to write comments on rough drafts or final drafts of essays, the data reflected essays in several stages of writing. Of the eight participants, four wrote comments electronically and four wrote comments traditionally. The instructors wrote comments in the mode they had planned to prior to inclusion in the study because I did not want to potentially introduce a bias before getting
good baseline data. All of these essays had personally-identifying information removed and were assigned numbers for ease of tracking.

The Data Cometh

In the end, 93 essays were collected from the eight participants. Four participants provided electronic comments, and four provided traditional comments. Of the essays themselves, 44 were electronically-commented, and 49 were traditionally-commented. Carol collected the fewest essays from her classes, having only received eight essays usable for the study. Still, this number represented a large enough sample of work that I felt that the work adequately represented her commenting style and characteristics.

Alice (another researcher interested in investigating these data) and I decided to use a coding rubric from Kwangsu Cho et al.’s 2006 study of essay comments in order to categorize the data. This study identified five types of comments: Directive, Non-Directive, Praise, Criticism, Summary, and Off-Topic. For the purposes of our study, however, we chose to essentially eliminate the summary category because summary comments were rare, and in the essays we received these comments overlapped other categories. A comment that might summarize what an author had just said never "merely" summarized. Rather, they were used to provide some kind of context for the overall comment. For example, a comment in Essay 190 said, "Wait: a couple paragraphs ago you said that very few women were working during Woolf’s day -- you can't have it both ways." The summary parts of these comments were inseparable from the rest of the comments, and in any case did not match the spirit of the category (presumably summary
comments are meant to show the writer a factual account of what the reader has leaned as s/he reads so the writer can tell if his/her meaning has gone astray), so due to rareness and ease of coding this category was eliminated from this study.

It may be important to note that Cho et al.'s use of the term "Directive" (as opposed to Non-Directive comments) is not be the same as Knoblauch and Brannon's Directive and Facilitative commentary (qtd in Straub 156-158), despite. Cho uses these terms to refer to the comment's personalization to the essay at hand rather than the level of appropriation that the instructor exerts over the text. That is to say, whereas a Directive comment under Knoblauch and Brannon might refer to an instructor telling a student to add a comma in order to fix a grammar error, under this study's coding rubric such a comment would have been classified as Non-Directive because the comment is not directed to the current essay. A comma error is an error of grammar, and grammar is universal across all essays. A grammar error is not an error only in the context of the essay or the student's intent. Rather, the error would be incorrect regardless of the writer or the content of the essay. Directive comments in this study refer to comments that somehow relate to the context, context, message, meaning, or communicative effectiveness of the essay.

Having a coding rubric that measured personalization to essays was helpful for this study because it offered a way to begin quantifying the level to which instructors were "connected" to the essays that they read. Anecdotal accounts of reading books often contain claims that reading on a paper makes readers feel "closer" to the writing. As well, this rubric had the potential to help determine whether the comments were substantial in the sense of being genuinely helpful for student revision. Recording merely the number
and word-length of each comment would have left the door open to unfairly evaluating a large number of superficial comments or grammar-based comments. Coding the data according to this rubric allowed me to have a quick and easy way to determine how many of the comments were geared towards meaning and revision, rather than style conventions. All marginal comments in the essays were coded by word-size and category, according to the rubric. We also recorded the total number of comments, as well as the length of the terminal comment if the instructor chose to write one.

Initial Coding Problems

Coding the data for this study ended up proving much more complicated than I had originally envisioned. The task seemed relatively straightforward at first. Word counts and comment counts would be time-consuming but simple. I expected that categorizing comments would provide a few tricky moments when instructors would inevitably write comments that could fall into multiple categories. Yet looking at the entire set of data collected for this study, the variety in the essays astounded me. While most essays fell in the de facto "standard" first-year composition essay of 4-6 pages, MLA style, there were many variants. I suspect that some of these variants were the result of the students' initiative, perhaps even against the instructors' wishes. Some essays included cover pages. Others were written in non-standard fonts, or with non-standard spacing and/or margins. A few included illustrations. Accommodating these various idiosyncrasies, and compensating for textual features and problems I had not envisioned when I began the study, required much adaptation of the study's core methodology.
Even from reading a few initial terminal comments, it became obvious that coding these comments would be extremely difficult. Even a relatively short terminal comment could contain multiple categories of comments, especially mixing Praise, Directive, and Criticism comments. Added to this, lengthy terminal comments would include multiple comments of each category. Thus, coding the terminal comments would have required breaking each terminal comment into "idea units" that could be categorized individually.

While this process was theoretically possible, the practical reality of coding so many papers according to such a complicated system would have required more time than was available, yet still had the potential to provide little meaningful information. Marginal comments, by nature of their large number and relatively short length, can be helpful for students by providing a way to induce rules from particular comments. While any individual comment might be fairly limited in scope, perhaps targeting only one sentence or a paragraph, the combined effect of many small/simple marginal comments can still be beneficial.

Terminal comments, on the other hand, are an instructor's chance to give students broad advice. I use terminal comments in my composition classroom as a chance to raise points about essays' overall structure, or patterns of error. Instructors interviewed later in the second phase of the study also reported using terminal comments in such ways. Overall, this creates an effect where the exact wording and tone may be important for students, as the greater length of the comments makes more nuanced interpretation and subtler shades of meaning possible. The rubric, being fairly coarse and only dividing comments into five categories, could not hope to adequately convey the complexity of the comments. Terminal comments represent an area where this study's quantitative methods
fell somewhat short of the best analysis for the data. Given these limitations, terminal comments were coded for word-count length alone. No categorical data was collected.

Other problems in data analysis soon became apparent as we analyzed the data. Essays with traditional comments often used the full margin space. This meant that some instructors wrote comments that were almost flush with the side of the paper for many comments. Photocopiers cut off the very edges of the pages. For electronic essays this was not a problem, as the word processors either still kept a small margin around the very edges, or we were able to obtain electronic copies of the essays to print fresh (thus avoiding the photocopier's inability to scan extreme edges). In most cases, the first letter or two or a word was cut off. Unfortunately, however, the wide variety of comments and instructors' ingenuity in finding ways to write large comments into small margins resulted in some essays having comments written vertically up the side of the page, so that the top or bottom halves of words were cut off. Others were simply written in such small handwriting that too much of the words were obscured to read.

In the cases where enough of the text was legible that I thought I could make an educated guess about the meaning, I coded these according to my best guess. In cases where I felt that the meaning of the comment would have been in question, I did not code the comment at all. Nor were those comments counted for word length or category because doing so would have led to sums of total comments and percentage breakdowns of categories that did not match across all categories. This may have skewed the data for traditional comments slightly, because the number of excluded comments is not reflected in the comment counts.
Instructors’ preferences for marking essays also introduced certain problems. While all instructors made marginal comments, some simply added editing marks (such as strikethroughs to stand for removing text, double-underlining to indicate capitalization, etc) to indicate corrections. These symbols had no good equivalent in length. Added to this, many instructors chose to strike through student text and write in direct replacement text, rather than suggest a change in the margins. Both traditional and electronic essays exhibited this tendency, as Microsoft Word's change-tracking feature allows instructors to make direct changes inside an essay, with those changes specifically marked as another writer's, pending approval from the original writer. I was thus able to tell which direct corrections were made by instructors in both styles of commenting.

These editing marks and direct word replacements were not coded during this phase of the study because the editing marks did not reflect comments by the teacher per se. Rather, they represented a strong controlling element in the essay that left the students with no choice. That is to say, these marks represented commands rather than comments and therefore fell somewhat outside the scope of the research.

Finally, each essay was given a page length. Counting the number of pages turned out to be one of the more difficult aspects to standardize over the course of the study. On one hand, counting only the number of comments per essay would have been problematic because the essays varied in length considerably. Some of the rough drafts were very short indeed, occasionally only 2-3 pages. A few essays were quite long indeed, including some that reached 15-20 pages. Counting raw instances would have advantaged instructors who donated longer final drafts to the study, or to those who simply assigned longer essays to begin with.
To compensate for this problem, I chose to include a page count with each essay. Doing so would allow me to calculate averages of comments per page that would negate any perceived advantage of lengthier essays.

This solution did have its share of problems as well. While an essay might appear homogeneous at first glance, with comments roughly evenly spread out throughout the essay, bibliographic pages soon turned out to be problematic. Some instructors chose to mark each stylistic error on these pages, and with the average APA and MLA citation requiring much punctuation, this had the potential to invite many marks indeed. As such, a Works Cited page could be the source of many Non-Directive or grammar marks, far out of proportion to the number of words on the page compared to a body page within the essay proper. So, too, were there problems estimating the number of words in cases of illustrations (some of which had captions, others not) which instructors occasionally commented upon.

In the end, I decided that the most general solution would simply be to count all pages, including bibliographic pages but not cover pages because instructors rarely wrote comments on those pages and the information on those pages (and the relevance to the essay as a whole) was very low, and in any case cover pages were quite rare amongst the essays (less than 2% of essays had cover pages). I counted partial pages as a final full page (thus essentially rounding up in all cases), also counting an illustration page as a full page because instructors often wrote comments about those illustrations. For the most part this worked well. Overwhelmingly, the essays used Times New Roman font in 11 or 12-point size. The others used fonts such as Ariel or Calibri, which would not have altered the page count unduly. Even an egregiously large font change such as Courier
New to Times New Roman would have only altered the page count by about 15-20%, or one page in most essays.

The "standard page" for this study became defined as double-spaced Times New Roman font, 12-pt, with 1" margins on all sides. Non-standard spacing such as single-spacing or one-and-a-half spacing was also easily converted to an equivalent in standard pages. Rounding up partial pages may have introduced some skew to the data, but no bias should have been introduced because both electronically and traditionally-commented essays were coded with the same page counts, thus skewing all data in the same way. Because this study aimed to study the difference between the two commenting styles, having the data skewed in the same way would still allow comparison of those data.

Other Small Sources of Error

While coding data, I had to make numerous minor decisions about how to code data. Many were not controversial per se, but rather a decision on how to treat certain cases so that the data collection would be consistent across essays. While none of these is likely to have caused a significant error bias, I will list the choices made here in case others would like to replicate this study or use it as a baseline for other research.

**Compound Words** were generally treated on a case-by-case basis. Multi-word adjectival phrases (such as "one-word") counted each constituent word. Compound words that essentially represented one unit of meaning (one morpheme) such as "fairground" were counted as one word.
Numbers represented an odd situation. I counted strings of numerals such as "22" as one word even though I counted written-out forms such as "twenty-two" as two words. I imagined an argument against this practice, given that the numerical form and the written form of a number should be identical in word-count. Conceivably, I may have skewed the data slightly if some instructors were more prone to write out numbers. But even if such an effect were present, I am confident that the influence on the final data tallies would be negligible because numbers made up such a small portion of instructor comments, and the difference in word count was small compared to the word counts of entire comments.

Results

Results in the first phase of the study showed relatively large differences between the two commenting styles. Electronically-commented essays generated a mean comments per page of 3.11 (N=44, standard deviation = 1.26), whereas traditionally-commented essays had a mean of 1.99 comments per page (N=49, standard deviation = 1.3). The difference was tested for statistically significant difference, and was found to be significant at the p<.02 level.

Electronic comments also tended to be longer in length. Mean word-length per comment was measured at 12.85 words (standard deviation = 5.85). Traditional comments had an average of 7.99 words (standard deviation = 4.99). This difference also tested significant at the p<.02 level. This value was calculated by taking the average of words per page and dividing by comments per page for each essay, then calculating the
average length based on those approximations, thus actually taking the average of averages. Such a measurement is less accurate than measuring each comment individually and calculating means, but doing so would have required keeping tallies of every comment in the study. The time involved would have been enormous, and given that the study aimed to measure relative differences precise accuracy wasn't crucial.

Both of these differences showed a much higher standard deviation compared to the means, indicating a higher variance in those comments. This indicates that the instructors who hand-wrote comments tended to write comments with a greater degree of varying lengths (shorter short comments, and longer long comments) compared to the more uniform electronic comments. Perhaps even more surprising, the number of comments generated per page was also more consistent amongst the electronically-commented essays compared to the traditionally-commented essays. The reason for the overall smaller variance for electronic comments is not clear, but the difference is striking.

If more comments and longer comments are taken as superior, then the data clearly favored electronic commenting. These essays provided students with 50% more comments per page, on average, as well as approximately 50% more words per comment. The implication would be more comments, and more detail in those comments. Percentage breakdowns of the comment categories also supported such a view. 61.7% of electronic comments were coded as Directive (D), compared to 37.6% of traditional comments. Electronic commenting resulted in 24% Non-Directive (ND) comments. Traditional commented resulted in 40% ND comments.
These figures are particularly interesting because they show that the electronically-commented essays were targeted content-specific issues more often than the traditionally-commented essays. The symmetry between higher D comments in electronically-commented essays and fewer ND comments in those essays reinforces this view.

Praise comments reversed the trend somewhat. 3.3% of electronic comments were coded as Praise (P), compared to 9.7% of traditional comments. Interestingly enough, one might expect Criticism (C) to appear more frequently in electronic comments, and those differences were found. Electronic comments resulted in 8.17% of Criticism, while traditional comments had 3.29% C comments. However, differences between C comments were not shown to be significant at the p<.05 level. Nor were Off-Topic (OT) comments statistically significant. The relative rareness of those comments (less than 10% of total comments each) made differences difficult to detect.

Initial Discussion

I was surprised by the magnitude of the findings. To be perfectly fair, I’d suspected that electronic commenting would produce longer comments, and perhaps slightly more. In my general experience, typing is much less physically-tiring than handwriting comments. For the population we were examining, this might have been doubly important. The first-year graduate assistants at Youngstown State spend the Fall semester learning composition theory, and then they teach two sections of Writing I in the Spring. Thus, most of the study participants had significant grading responsibilities in addition to
meeting those responsibilities for the first time. I'd personally found that I tended to grade in long sessions, sometimes grading essays in stretches of 2-3 hours at a time. While this isn't an extremely long amount of time, writing comments for essays can quickly tire out an instructor's hands.

The physical limitations of paper also suggested why traditionally-commented essays did not provide comments of the same length (or number) as electronically-commented essays. As noted earlier, Microsoft Word does a very good job of fitting instructor comments into the available space. Given that the text in comment bubbles is fairly small (I would roughly guess they are recorded in an 8-point font), even the most verbose instructor's comments fit onto the page neatly. Traditional comments, on the other hand, always seemed on the brink of overloading the page. Even instructors with small handwriting tend to fill a one-inch margin in short order. Comments longer than a short description inevitably filled a few lines of text. Occasionally, a single comment might take up a quarter or a third of a page's margin. Presumably, if instructors were at all concerned with filling up all of the available white space, they would logically limit themselves to only the most important points to make, or try to keep comments down to minimal descriptions.

The full story of why the study observed longer and more numerous comments in electronic feedback probably wasn't limited to any one of these factors. It was probably a mixture of the two in different ratios per the personality and needs of each individual instructor. Indeed, there may have other factors in play that haven't been considered so far. But taken on the whole, I was still very satisfied with the initial findings.
If instructors could be influenced to personalize their feedback by using electronic methods of comment-writing, then the future of composition may unintentionally benefit simply as a result of spreading technology. As colleges move to e-portfolios to collect student work, and assignment collection via e-mail becomes more common, instructors may find themselves more likely to use word processors to write feedback. Additionally, these data indicate that instructors may not need to be overly concerned with the efficacy of instructor feedback in distance-learning courses. Although this study is not definitive by any means, the early indications are very positive indeed.
In Chapter One, I began researching teacher comments in the best way that I could: by examining the comments generated by graduate students at Youngstown State University. The results I'd found demonstrated clear trends, and electronic comments looked like they would benefit student writing far more than traditionally-written comments. Instructors tended to write about 50% more electronic comments, those comments tended to be approximately 50% longer, and best of all the electronic comments showed a significantly higher proportion of Directive comments. This last finding was the most telling of all because it indicated that electronic comments tended to address content issues in the essays. Going into the next phase of the research, which I discuss in Chapter Two, I was optimistic about the results that I'd already seen, and my primary goals going forward were to expand the study to begin accounting for some of the elements that I'd originally controlled.

The initial results were surprising and full of potential for future research. Knowing that Alice and I had found clear differences that had powerful implications for future teacher training fueled my desire to expand the research. Originally, I'd expected modest differences between the two commenting styles, or perhaps some kind of conflicting data that suggested pros and cons for both styles that might not lead to conclusive results. However, the data analysis for Phase One exceeded my wildest imagination. Indeed, it was difficult to restrain myself from telling every composition
instructor I knew that they should immediately begin commenting electronically to take advantage of the longer and more numerous comments.

The study also began to raise more questions than it had answered. Were there gender differences between men and women while writing comments? What differences might we see from more experienced instructors? Did the medium of commenting affect the tone of the comments? Why did electronic comments result in more Criticism comments? Did this imply that instructors working through a computer were somehow "distanced" from their students and therefore more likely to be critical? If so, then why would the comments show more Directive comments? Were there ways to combine the two forms of commenting, perhaps through graphics tablet? And perhaps most important of all, what did students think about traditional and electronic comments?

It soon became obvious that tracking down all of these research possibilities would be impossible, as much as because the project would encompass too wide a scope as well as time limitations. While Alice wanted to investigate the discourse of instructor comments (an understandable goal, given that we'd assembled a fairly large corpus for such research), I was more interested in following the research along its previous lines. In my mind, that meant collecting more data and setting up new conditions for that data gathering.

Phase One of the research had worked well as an initial step, but many of the study's characteristics that made it a well-controlled study also limited its ability to draw definitive conclusions. The study still had a population drawn almost exclusively from graduate assistants who were teaching for their first semesters. Beyond that, Phase One was only able to generate correlational data. That is to say, Alice and I collected data
from instructors who commented in the mode of their choice. A strictly-controlled study would have randomly selected graduate assistants from the total pool, then randomly assigned those instructors a commenting mode. This random sampling would eliminate any potential biases. It would have been an ideal methodology.

Life, it turns out, is not ideal. I did not have access to the total pool of graduate assistants. I had access to those who were generous enough to volunteer for my study, and this introduced practical limitations. By allowing instructors to commenting in the method of their choosing, I eliminated any "training bias" introduced by switching instructors over to a new commenting method. This worked well to give a true representation of what electronic commenting might accomplish, as otherwise low word counts in electronic commenting might have been attributable to instructors becoming frustrating with typing or difficulties with the software. It also meant that instructors who wanted to comment electronically did not have to deal with hand cramps during extended grading sessions.

Unfortunately, it also meant that I ran the risk that instructors who are personally inclined to write more comments, or longer comments, might naturally be drawn to electronic commenting. Or alternately, there might be something about instructors who write numerous Praise comments that draws them to hand-write feedback for student essays.

However, even creating a well-controlled study wouldn't necessarily provide the data I needed to make claims about different feedback methodologies causing different comments to be written. In order to do that I'd have to go still one step further and have a portion of my control element switch commenting methods, then compare to the original
data or (preferably) a continuation study of the remaining control element who did not switch commenting modes but did continue to be observed by the study.

Searching For a Better Methodology

I began my expanded study by first contacting a few full-time faculty members to take part in the study. This would serve a dual purpose: first, it would allow the study to gather more data. With a pool of only eight participating instructors, I still had a limited enough data set that any instructor willing to contribute essays would have been helpful. Second, it would allow me to start adjusting the data for the experience bias potentially present in collecting data from only graduate assistants teaching their first semesters. The full-time faculty teaching Writing I sections at Youngstown State University in were asked to participate in the study. The vast majority (in excess of 90%) of sections at Youngstown State University during the Fall and Spring terms are taught by graduate assistants and adjunct faculty, and would therefore be the logical population to study, but I instead focused on the full-time faculty in order to remove any bias that may have occurred due to educational training as well as experience.

Taking the next step, carrying out Phase Two of the study with a controlled experimental methodology, proved nearly impossible. While some instructors had been willing to participate in Phase One of the study, commenting in their preferred method, few were willing to even consider the possibility of using another commentary style. For my part, I did my best to make participation as easy and un-intimidating as possible. I contacted all PhDs teaching Writing I at Youngstown State University in the Fall 2009
semester, as well as adjunct faculty members. In total, nearly 20 composition instructors were approached to take part in the study.

I explained the purpose of the study and its brief history to each instructor. I then explained that I was looking for volunteers to switch their usual commenting style, either from traditional to electronic or vice-versa. It turned out that of the instructors invited to participate, all preferred to hand-write comments. I then informed each that I would be happy to provide either one-on-one training in Microsoft Word's commenting features, or I would be equally happy to lead a group in a workshop if that would be preferable.

I knew that finding instructors to take part in the study would be difficult. In Phase One, there had been some resistance to participation in the study. Many instructors were concerned that participation would require too much time, and a few feared that the analysis would somehow be used to show a somehow flawed teaching practice. The latter was an understandable fear, given that the instructors were generally teaching for their first semester. Confidence issues were to be expected.

Fears of time commitments were more difficult to assuage, but Alice and I assured instructors that their students would find the process entirely transparent (as the study did not ask the students to take any action they would not have normally taken, and nor did it ask them to take those actions differently). The instructors themselves only needed to make essays available for reproduction. In the case of electronically-commented essays, that generally meant e-mailing copies of the essays to either me or Alice. In the case of traditionally-commented essays, that meant providing the essays so that Alice and I could make photocopies of the essays. In those cases, I was generally able to return essays in under an hour. Even in the worst cases, students received their
essays one class after they normally would have received the essays (two days), and even this was only the case for final drafts.

As far as instructor time was concerned, volunteers were assured that participation in the study would not require undue amounts of time. They would grade essays fundamentally the same as they always had, and I would collect the essays from them as well as return the essays. Alice and I essentially made the project as close to zero-effort as possible.

With Phase Two of this research, such assurances were no longer possible. Instructor enthusiasm for the research, predictably, withered as well. While all instructors were familiar with the idea of electronic commenting, and some had attempted using these methods at some point in the past, none would consent to be randomly assigned a commenting style. Instructors who had not used electronic commenting were not interested in receiving training.

What amazed me most about the refusals to use electronic commenting was the sheer variety of reasons given. One instructor said he disliked staring at computer screens for the lengths of time required to grade essays. Others stated that writing comments for essays handed in on paper copies was more convenient because the instructors would not need to have a computer, and so could read essays on-the-go during a few free minutes here and there.

On top of those reasons, I heard almost all of the reasons I expected to hear, such as not feeling comfortable with the software. Several instructors remarked that they did not want to grade essays electronically because doing so would require collecting the essays via e-mail or some other electronic filing system (such as the ubiquitous WebCT,
Blackboard, or e-mail). Adding this dimension does add some complications to collecting student work. Although in an ideal world electronic essay collection would in fact be faster and easier, the truth is that many of Youngstown State's students are not well-familiar with the university's e-mail and WebCT system, which is based on software that is showing its age. Various user-interface quirks such as requiring multiple button-presses and navigation through multiple screens (including scrolling to the bottom of a page to click buttons to verify actions) occasionally confuse students who are used to more streamlined interfaces. Even generally computer-literate students (and occasionally instructors) can send e-mails without attachments after forgetting that they must "submit" after "accepting" or "attach" after "browsing."

Added to this, the ecology of software available makes electronic submissions somewhat more complicated than it might appear. Though Microsoft Word is standard in all of Youngstown State's computer labs, it's sold with an educational discount through the university, and the vast majority (everybody I contacted while conducting this research) of instructors, the fact still remains that file format compatibility makes electronic file collection difficult. Even adopting Microsoft Word as the standard has not entirely solved this problem. When the 2007 version of Word was released, Microsoft introduced a new .docx file format to replace the .doc format. Older versions of Word (which are still common among students) can read the .docx files, but in order to do so the user must download a file compatibility update from Microsoft's website. The file is free and available to all users of Word, and is relatively easy to install, but most students (and indeed, many instructors) simply do not know that such tools exist.
If that weren't enough Microsoft also offers an office productivity software suite named Microsoft Works. This software is often pre-installed on computers offered through many retail channels for free, so students often use Works instead of purchasing Word. Unfortunately, Microsoft Works uses the .wps file format, which again requires a compatibility download to use in Microsoft Word. At that point, the single-letter difference between the software titles (which predictably causes all-too-frequent miscommunications) is just insult on top of injury. Even in a world with seemingly universal standards for word processors, the fact of the matter is that the few incompatibilities that do exist seem to stymie instructors even more than ever because instructors lack the powerful computer literacies required to troubleshoot such problems when they overwhelmingly work in a world where standards are adhered to and computers "just work."

It quickly became obvious to me that the factors that play into instructor commentary styles are much more complicated and nuanced than even a composition instructor might expect. But interestingly enough, even these concerns paled before the time constraints that instructors must work within. Almost every instructor who declined participation in the study cited two reasons: First, fears of not having sufficient time to grade or write feedback for essays. The general picture of life as I saw it through the instructors' eyes was one of amazing precision. Instructors had exactly enough time to complete all of the work they had. Even one more task, or slightly more time taken with any task, would lead to something not getting done. Instructors had no "wiggle room" in their schedules whatsoever.
Second, instructors feared that changing their commentary style would somehow produce comments that were different or unfavorable. The instructors I spoke to were not specific about these points, but a few mentioned not wanting to switch styles because of upcoming performance reviews. I didn't press instructors for details of why they did not wish to participate (I merely offered any assistance I could provide to encourage participation), but the message was clear: instructors who had found an acceptable method that worked had too much to lose by adopting a new technique, even if that new technique might offer pedagogical benefits.

Ultimately, the final answer was clear: instructors adamantly refused to change their commenting on command, even if it was in the name of academic research. Many expressed regret, and assured me that they would participate if the situation were different. If they didn't have so much other work to complete (so many essays to grade in general, other commitments to the university, or other personal responsibilities), they would have gladly helped. But what I was asking was simply not feasible.

I attempted to negotiate as best I could with the instructors. I asked them to comment only one section for the study (keeping in mind that I asked for data from only one essay during the semester). Overwhelmingly, I was told that changing commenting styles for even one essay out of the semester would throw schedules off too much. I considered asking for a smaller sample, perhaps half of a section, but I feared that less than one section would not have provided enough data from each instructor to form an adequate profile of his or her feedback style.

In the end, only four instructors (including myself) participated in the experimental portion of Phase Two. These were Bill, Fran, Nancy, and Leon. All four
instructors had already used electronic commenting in the past, and were comfortable with the system. Bill and Fran had participated in Phase One of the study, and both volunteered to provide a set of essays with comments in the opposite style for Phase Two. Nancy and Leon were participating in the study for the first time, and they volunteered to provide two sets of essays: one traditionally-commented and the other electronically-commented.

With only four instructors, randomly assigning instructors to a style of commenting was pointless. The data set would be too small to gather statistically significant data, and gathering enough participants to create such a study was simply impossible. I had already approached everybody in the Youngstown State University English Department, and one friend who worked at another university, who I had a personal connection with. All of these people were people I was friends with, had worked with in the past, or were generally known in the department to support research. If asking almost 20 people only resulted in three volunteers (me being one of the four to participate) then asking the remaining 30-40 instructors (many of whom I had never met and would not have recognized me by name) in the English Department would have only garnered perhaps 1-2 volunteers even in a fairly optimistic scenario.

As much as I wanted all the data I could gather, I knew that going from four participants to six participants would not change the study radically. I would still be left with a fairly small set of data, and taking the time to contact those remaining instructors (finding them during office hours, attempting to convince them to take part in the study, etc.) would require considerable time. Every day that passed while I missed people at office hours, or failed to talk somebody into participating, meant that I was that one more
day into the semester. Soon enough, first essays were being assigned and collected. Finally, and with some regret, I had to accept that the study wouldn't have the scale of participation I wanted.

Collecting More Data From Fewer Subjects

Partially to compensate for the reduced data set, I took measures to expand the data collection. Thought the study might have included a relatively small number of participants, I decided that it would be a good idea to gather as much data as possible. This might help mitigate the breadth of the data by providing some depth. In addition, I hoped that additional data collection would help to answer some of the questions posed in Phase One. For example, which commentary style did students prefer? Or why did instructors choose their respective styles? Gathering information to give a fuller context of the situations and motivations that writing feedback takes place in might eventually be key to understanding the reasons why I'd seen the effects observed in Phase One.

To this end, I created surveys for the students to fill out that asked them about their opinions and preferences when reading instructor feedback. A simple and straightforward Likert scale survey appeared to be the best fit for the participant pool. Between my instructors, eight sections of Writing I would be asked to participate in the study. Each section had a theoretical maximum of 25 students each, though I expected that between absenteeism and students declining to participate, only around 15 from each section would in fact volunteer. But even that rate of participation would have left over 150 students to offer data. The survey was composed of 30 questions, one of which was
designed to track the commentary style the student observed and the other 29 designed to
examine student preferences and habits while reading feedback and revising essays. I
limited the surveys to 30 questions to keep the size of the survey minimal. With so few
questions, combined with a straightforward Likert scale answer scheme, the survey could
be administered.

In addition to this survey data, I also invited each student participant to take part
in a short interview about their preferences for instructor feedback. Students were also
invited to write short accounts of their experiences with reading feedback. I had some
concerns that students would not be inclined to write about how they feel about instructor
feedback. The triangulation offered by this approach was (at least theoretically)
considerable. Where survey data would do well to canvass a large number of students and
provide broad trends, the interview data would provide depth and context. If the two
matched, I could reasonably make conclusions about motivation as well as action.

I interviewed three instructors. I had originally planned to survey the instructors
in roughly the same manner as the students, save with an emphasis on writing comments
rather than reading them, but ultimately the surveys were dropped from the study. With
only six instructors taking part in Phase Two, the breadth of data necessary to make
quantitative data collection would have been impossible to gather. Any results I gathered
from the survey would have represented too small a sampling to generalize to the greater
population of instructors, and the variability of the data would have made statistical
significance impossible to determine except in the clearest cases of unanimity, which
would have been reflected in the interview data as well.
Adding More Analytical Methods: Editing Mark Categorization

After analyzing the Phase One data, I became interested in tracking the so-called "editing marks" so frequently found in traditional comments. I'd noticed these marks on virtually all of the paper-copy essays, and it seemed that much of the "work" of the electronic comments had been duplicated in the editing marks. Indeed, if my conjecture that the physical space available for comments played a role in why electronic comments generated so many more words, then it would only have been natural that instructors would have found ways to efficiently use the space that was available. Editing marks would have been quick, require relatively few pen-strokes, thus being less tiring to write than writing the same information out in words. It was also worth noting that some of the instructors who wrote electronic comments also used unclassifiable marks. Some used Microsoft Word's highlighting feature to indicate sentences that required some form of attention (usually a revision to correct a grammar error, but occasionally a stylistic error such as word repetition).

Taken together, these forms of feedback represented an important part of instructors' strategies for feedback. While I'd made the choice to exclude those data from analysis during Phase One, when I began Phase Two, I went back and examined the original data set to find ways that those features could be worked into the data analysis. I decided to code those marks as "grammar marks" because most of them were meant to convey some kind of grammar point, such as a comma error.

Coding grammar marks was relatively straightforward most of the time, but in a few cases I decided to code multiple changes to an essay as one single mark. These
examples were places where I felt that the instructor had fundamentally made one mark because several points hinged upon the same principle. For example, if an instructor corrected a run-on sentence by inserting a period, then capitalizing the word that would begin the new daughter sentence and inserting the period were coded as a single grammar mark because the marks could be taken as a system of corrections that stemmed from one central idea. Therefore, the grammar marks represented conceptual corrections rather than the expression of those corrections.

The Data Cometh, Part the Second

In total, Phase Two of the study added six instructors to the study (Ian, James, Kelly, Leon, Mike, and Nancy). Mike and Nancy were all PhDs in Composition. Ian was PhD in Composition and Education. James and Leon were MA in English, though both were full-time instructors at Youngstown State who regularly taught courses in the composition sequence, as well as other writing courses such as technical writing. Kelly was a graduate student.

Over the duration of the entire study, there were 14 instructors total. The study collected 255 essays in total. Of these, a few were rendered illegible due to problems with a photocopier. Another few were accidentally copied twice. After removing these anomalies, the study collected 247 essays in total. Of these, 106 were commented electronically. 141 essays were commented traditionally. Four instructors (Alice, Carol, Helen, and James) provided only electronic comments. Six instructors provided only
traditional comments (Daphne, Edward, George, Ian, Kelly, and Mike). Four instructors provided both styles of comments (Bill, Fran, Leon, and Nancy).

My first step in analyzing the data was to create a new baseline made up of all of the data that I’d collected over Phase One and Phase Two. Doing so would allow me to compare the broadest overall trends of electronic and traditional comments. I then planned to create smaller "commenting profiles" made up of the data subset I'd collected.

From the grand data set, electronic commenting generated a mean words per page of 32.04, with a standard deviation of 24.59. Traditional comments generated 21.78 words per page, with a standard deviation of 11.08. This difference tested statistically different at the p < .02 level.

Electronic comments generated a mean comments per page of 2.53, standard deviation of 1.25. Traditional comments generated 2.90 mean comments per page, standard deviation of 1.55. This difference also tested statistically significant, though only at the p < .05 level.

Immediately, the picture of the grand data set had started to look strange. The number of comments per page had reversed the trend from the Phase One data, which had shown a mean of 3.11 comments/page for electronic comments and 1.99 comments/page for traditional comments. Not only were the numbers different, but the traditionally-commented essays now generated more comments.

A rough calculation of mean comment length calculated by dividing the words/page by comments/page showed that electronic comments generated an average comment length of 11.86 words, standard deviation of 6.60. Average traditional comment length was roughly calculated at 8.02, standard deviation of 4.11. These numbers were
not entirely surprising, given that the relative difference between the two was about the same as measured in Phase One.

Perhaps not surprisingly, instructors tended to create more grammar marks on traditionally-commented essays (mean of 3.56 grammar marks per page, standard deviation of 6.76) compared to electronically-commented essays (mean of 1.65 grammar marks per page, standard deviation of 2.39). This difference tested statistically significant at the p < .02 level.

Calculating mean terminal comment length wasn't quite as easy as in Phase One. Several essays did not include terminal comments, either because the instructor had not written one (as might be in the case of a rough draft) or the terminal comment had been replaced by smaller comments on a grading rubric sheet. Counting those essays as a terminal comment with a length of 0 (zero) would have been quick and simple, but it risked pulling down means tremendously and would have added a huge amount of variance to the data which would have skewed the standard deviations. Instead, for these calculations I chose to exclude essays with no terminal comment. With that in mind, electronically-commented essays showed a mean terminal comment length of 104.87, standard deviation of 74.86. Traditionally-commented essays showed a mean length of 120.52, with a standard deviation of 79.03. Surprisingly, this difference did not test statistically significant, even at the p < .05 level.

Even though the difference in terminal comments wasn't found statistically significant, I can't help but wonder about why the trend would have reversed itself in the way that the study observed. Seeing longer comments from electronic feedback in general isn't surprising. As I noted earlier, there are numerous physical limitations on
hand-written comments. But one would assume that these limitations would apply to terminal comments as well as marginal comments. If I'd observed slightly longer terminal comments from electronically-commented essays, that might not have been surprising. But it's shocking to see that in fact traditional comments generate longer terminal comments. Again, here it's tempting to read results from the data, but with a p value of approximately .15, any conclusions are too likely to have been the result of random coincidence. Regardless, the lack of clear results is troubling in and of itself.

The next step was to examine the frequency of Directive (D), Non-Directive (ND), Praise (P), Criticism (C), and Off-Topic (OT) comments in the two groups of essays. Here, D comments made up 50.50% of electronic comments, with a standard deviation of 25.63%. Traditional comments showed 48.14% D comments, standard deviation of 26.05%. This difference did not test statistically significant at the p < .05 level.

ND comments showed a similar trend. 27.59% of electronic comments were coded ND, with a standard deviation of 21.55%. Traditional comments were composed of 29.58% ND comments, with a standard deviation of 25.60%. Not surprisingly, this difference did not test statistically significant either.

P comments made up 8.91% of electronic comments, standard deviation of 12.45%. These comments made up a somewhat greater 13.01% of traditional comments, with a standard deviation of 13.01%. Though this difference appears fairly large, the standard deviations are extremely large in comparison to the mean results. This large standard deviation indicates that the rate at which instructors write Praise comments
varies widely. Some may write many, but others write very few. As a result, this
difference was not statistically significant either.

C comments made up 5.34% of electronic comments, with a standard deviation of
8.90%, while traditional comments were composed of 4.15% of C comments, with a
standard deviation of 8.30%. Here the standard deviations are extremely large compared
to the means, and this difference did not test statistically significant either.

I suppose that it's as much for completeness as anything else that I record the OT
figures. Electronic comments showed 5.78% OT comments with a standard deviation of
10.89%. Traditional comments showed 4.93% OT comments, with a standard deviation
of 11.00%. As I'm sure a careful reader would guess, this difference wasn't statistically
significant either.

To say that I was surprised would be an understatement. I was surprised,
mortified, dismayed, confused, and angry. With the exception of the lengthier electronic
marginal comments, all of the effects I'd seen in Phase One had disappeared. Not only
that, but differences had largely disappeared. Now it looked like there were no
differences between the two commenting styles whatsoever.

Searching for Answers

I undertook a few more steps to determine exactly why the Phase One results had
disappeared. The most likely bias to have been introduced into the study was that of
experience. Phase Two had collected data from a much wider range of instructors, from
PhDs to full-time instructors with MAs. Splitting those instructors into groups to create
"experience profiles" might have demonstrated why some effects would have been confounded. This was to be expected, and Cho et al. had found different rates of Directive, Praise, and Criticism commenting in a study of feedback that included several groups divided into experience categories (274-280).

Dividing my own participants into experience profiles, however, painted a rather disjoint and frustrating picture. The graduate assistants wrote an average of 2.99 comments per page across all essays in the study. The full-time MA instructors wrote 3.22 comments per page. This shows the rise that we might expect from more experienced instructors, yet the number drops dramatically for PhDs, who only wrote 2.17 comments per page.

More comparisons of the graduate assistants versus the PhD instructors were similarly bizarre. While the graduate assistants made an average 3.26 grammar mistakes per page and wrote average terminal comments of 107.61 words, the PhDs made an average of 1.51 grammar marks per page, and wrote terminal comments of 47.44 words on average. Perhaps complicating matters somewhat more, the full-time MA instructors wrote an average of 1.09 grammar marks per page and an average terminal comment of 123.70 words.

<table>
<thead>
<tr>
<th></th>
<th>Grammar marks per page</th>
<th>Comments per page</th>
<th>Terminal Comment Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA</td>
<td>3.26</td>
<td>2.99</td>
<td>107.61</td>
</tr>
<tr>
<td>MA</td>
<td>1.09</td>
<td>3.22</td>
<td>123.70</td>
</tr>
<tr>
<td>PhD</td>
<td>1.51</td>
<td>2.17</td>
<td>47.44</td>
</tr>
</tbody>
</table>
Excepting the terminal comments, the general picture is that of falling comment numbers and terminal comment length, yet in both cases the trend is violated by the full-time MA instructors. The differences between graduate assistant and PhD are statistically significant at the P < .05 level, but it's not clear what to make of them. While these data could be used to support an argument that instructors tend to comment less as they gain experience, it's equally likely that the results are simply the result of a sampling bias.

Though this study includes 145 essays from graduate assistants and 64 essays from PhDs, in the end the data is still drawn from a very small pool of instructors (eight graduate assistants and three PhDs). It is easy to aggregate the data and create profiles of each group, but in the end this aggregation can make the data look more representative than it truly is. A more accurate method of comparing the two groups would be to calculate averages for each instructor, then calculate profiles of the groups from this data (thus taking averages of averages) but with the low number of participants in this study such a procedure would be pointless because the extremely limited data would make detect differences between the most egregious data patterns.

Naturally, the next thing I did was try to find answers by analyzing the data in other ways. I isolated all of the data collected from Bill, Fran, Nancy, and Leo (my four participants who'd provided both styles of commenting). If the results from Phase One had been valid, then I'd almost certain see evidence of those differences in these four participants as they moved from one commenting style to the other. I've compiled those data in the table below.

<table>
<thead>
<tr>
<th>* Indicates a statistically significant difference (p &lt;</th>
<th>Electronic</th>
<th>Traditional</th>
</tr>
</thead>
</table>

43
These were not the results I expected to say the least. I'd gone from finding more comments, and longer comments to just longer comments. Even more surprising, the original trend of more Directive comments for electronic feedback had been reversed in favor of more Directive comments for traditionally-commented essays. Nearly all of the other effects I'd noticed in Phase One (more ND comments in traditional essays, more P comments in traditional essays, more C comments in electronic essays) had disappeared.

I still wasn't sure that I understood what the data were telling me, so I did the only logical thing: I ran more statistics and sliced the data even finer. I pulled data for each of the four experimental participants and examined their individual differences. I've since created tables to summarize each instructor profile.

In each of the tables below, one asterisk (*) indicates statistical significant at the $p < .05$ level, while two asterisks (**) indicates significance at the $p < .02$ level.
<table>
<thead>
<tr>
<th></th>
<th>%P comments**</th>
<th>%C comments**</th>
<th>%OT comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.30</td>
<td>13.24</td>
<td>2.94</td>
</tr>
<tr>
<td></td>
<td>13.38</td>
<td>3.21</td>
<td>0.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Fran</th>
<th>Electronic</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean comment length</td>
<td>5.98</td>
<td>5.68</td>
<td></td>
</tr>
<tr>
<td>Mean comments/page</td>
<td>2.28</td>
<td>3.57</td>
<td></td>
</tr>
<tr>
<td>Mean grammar/page**</td>
<td>0.00</td>
<td>2.14</td>
<td></td>
</tr>
<tr>
<td>Mean Terminal**</td>
<td>60.79</td>
<td>116.43</td>
<td></td>
</tr>
<tr>
<td>%D comments</td>
<td>38.64</td>
<td>26.55</td>
<td></td>
</tr>
<tr>
<td>%ND comments</td>
<td>29.43</td>
<td>46.66</td>
<td></td>
</tr>
<tr>
<td>%P comments</td>
<td>19.12</td>
<td>8.98</td>
<td></td>
</tr>
<tr>
<td>%C comments</td>
<td>14.96</td>
<td>3.12</td>
<td></td>
</tr>
<tr>
<td>%OT comments</td>
<td>3.11</td>
<td>2.86</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Nancy</th>
<th>Electronic</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean comment length</td>
<td>11.07</td>
<td>10.53</td>
<td></td>
</tr>
<tr>
<td>Mean comments/page**</td>
<td>1.59</td>
<td>2.71</td>
<td></td>
</tr>
<tr>
<td>Mean grammar/page</td>
<td>1.72</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>Mean Terminal**</td>
<td>7.23</td>
<td>29.65</td>
<td></td>
</tr>
<tr>
<td>%D comments**</td>
<td>56.55</td>
<td>75.25</td>
<td></td>
</tr>
<tr>
<td>%ND comments</td>
<td>17.31</td>
<td>7.50</td>
<td></td>
</tr>
<tr>
<td>%P comments</td>
<td>10.74</td>
<td>14.23</td>
<td></td>
</tr>
<tr>
<td>%C comments</td>
<td>0.55</td>
<td>1.51</td>
<td></td>
</tr>
<tr>
<td>%OT comments</td>
<td>7.16</td>
<td>3.12</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Leon</th>
<th>Electronic</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean comment length**</td>
<td>16.23</td>
<td>9.15</td>
<td></td>
</tr>
<tr>
<td>Mean comments/page</td>
<td>3.18</td>
<td>4.02</td>
<td></td>
</tr>
<tr>
<td>Mean grammar/page*</td>
<td>.89</td>
<td>3.31</td>
<td></td>
</tr>
<tr>
<td>Mean Terminal*</td>
<td>209.63</td>
<td>120.70</td>
<td></td>
</tr>
<tr>
<td>%D comments**</td>
<td>39.67</td>
<td>59.54</td>
<td></td>
</tr>
<tr>
<td>%ND comments**</td>
<td>53.37</td>
<td>26.20</td>
<td></td>
</tr>
<tr>
<td>%P comments</td>
<td>5.41</td>
<td>6.92</td>
<td></td>
</tr>
<tr>
<td>%C comments</td>
<td>0.96</td>
<td>1.10</td>
<td></td>
</tr>
<tr>
<td>%OT comments</td>
<td>1.59</td>
<td>5.04</td>
<td></td>
</tr>
</tbody>
</table>
It's hard to make much out of these tables. Of the four instructors, three show a significant difference in grammar marks per page. Two show a significant difference in mean comment length. Similarly, two instructors showed a significant difference in percentage of Directive comments. The only factor that all instructors showed a significant change in was terminal comment length, and even this is somewhat problematic as only three instructors show that the traditional commenting produces a longer terminal comment.

Taken on the whole, it's difficult to say that there are clear trends amongst these instructors. Certainly the Phase One observation that instructors tend to write more comments when providing electronic feedback was effectively nullified. In all four cases, instructors wrote more comments when writing feedback traditionally. It's possible that this small group of instructors is simply too small a group to adequately represent composition instructors as a whole, but data taken from all participants shows slightly more comments per page for traditional essays, so these data triangulate relatively well.

The earlier finding that instructors tend to write more Directive comments, with a corresponding drop in Non-Directive feedback, when writing electronic feedback was also effectively overturned. Rates of D comments in the total dataset were nearly identical (50.5% and 48.1% for electronic and traditional, respectively), and only two of the four experimental instructors wrote more D comments. The ND comments do not show the corresponding increase in the total dataset, and indeed the only instructor out of the experimental group to show a significant difference when switching modes
decreased ND comments in traditional feedback, rather than increase as was predicted from Phase One.

Due to the relative rarity of P, C, and OT comments, it's not surprising that finding trends is difficult. Standard deviations for those categories are extremely high, relative to the mean. Putting those figures into context is difficult. A standard deviation is a measure of variance in the data, but it could equally imply that each instructor has a unique style that widely differs from other instructors (while that instructor tends to write comments in the same general pattern or distribution) or it could mean that all instructors tend to write P, C, and OT comments sporadically. I've compiled the individual profiles for standard deviations for Bill, Fran, Nancy, and Leon in a table along with data from the entire data set to demonstrate what the data actually show.

<table>
<thead>
<tr>
<th></th>
<th>All instructors</th>
<th>Bill</th>
<th>Fran</th>
<th>Nancy</th>
<th>Leon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Praise (P)</td>
<td>126%</td>
<td>153%</td>
<td>83%</td>
<td>102%</td>
<td>130%</td>
</tr>
<tr>
<td>Criticism (C)</td>
<td>184%</td>
<td>128%</td>
<td>101%</td>
<td>628%</td>
<td>237%</td>
</tr>
<tr>
<td>Off-Topic (OT)</td>
<td>207%</td>
<td>346%</td>
<td>218%</td>
<td>146%</td>
<td>146%</td>
</tr>
</tbody>
</table>

This figure shows the ratio of standard deviation to the mean as a percent.

The figure above indicates that although some of the instructors did indeed have less variance in their comment-writing compared to the group, on occasion they were more variable than the group. Certainly, in these three classes of comments it seems that no instructor is ever truly predictable. Even the smallest standard deviation is 83% of the mean, which indicates fairly large variance in the observed essays. It's possible that the
standard deviations would lower somewhat if more data were collected, but it's not clear how much the standard deviations would fall.

Confounded Results

Even with these data, there are some confounding factors that call some of the results into question. While Fran and Leon both submitted final drafts with comments, Bill and Nancy submitted one set of rough drafts and one set of final drafts. It's easy to imagine that their pattern of commenting may not reflect the feedback style, but rather the kinds of comments that they write for different steps in the writing process. This same factors would have affected the grand aggregate data, but the increased number of participants would have helped to alleviate any bias.

Final Conclusions?

Looking at the data as a whole, the only conclusion that seems to hold is that instructors tend to write longer comments when generating comments electronically. However, we must temper this conclusion with the understanding that instructors who write traditional comments tend to write more comments per page. Electronically-commented essays tend to generate approximately 50% more words per page, but this greater word-count is largely due to each comment measuring about twice as long (the greater number of comments on traditionally-commented essays making up the difference in total word count per page).
Interpreting the numbers: False Traps

When looking at the numbers, the first thought that we have to be wary of is assuming that the high standard deviations observed in this study are cause for alarm. It might be tempting to say that instructors need to lower those values or stop putting comments "all over the place." I suppose that I could even imagine a world wherein instructors were encouraged to meet certain quotas of comment categories, perhaps 60% Directive comments, at least 15% Praise comments, and no more than 5% Criticism.

But while consistency might appeal to our souls and reassure us that "the system is working," I'm not convinced that any such determination can actually be made. We must be wary of reading too deeply into the numbers, because the data can only answer the questions we have asked. In this study, there was no attempt made to understand why various comments were made, or how they were made.

The instructors in this study may have shown high standard deviations in many categories, but it's entirely possible that they did so in order to customize the comments for the students. By the same token, it might be easy to look at the rates of Non-Directive comments and declare that instructors should find ways to make all comments either Directive or Praise. Again, at a glance this might be a laudable goal. Presumably we do want instructors' comments to be individualized to each students' needs. But the Non-Directive category encompasses many points, including style and writing convention. This category included advice about citation styles and essay formatting as well as grammar. While these concerns might not be the foremost writing concerns that students
should think about, ignoring them altogether is not an acceptable situation either. In this sense, lengthy Non-Directive comments would be preferable because it would indicate that the instructor is attempting to convey some important information about the grammatical structure in question. Perhaps the instructor is explaining the principles behind the revision, giving examples of ways to revise, or giving detailed instructions for how the student can locate the error in the future. Here, a Non-Directive comment is far preferable to a grammar mark that simply adds a comma or capitalizes a proper noun with no explanation.

Finally, we should also remember that most of the data presented in this study are percentages. Being divorced from the context of an essay can be especially dangerous here, because an otherwise poor pattern of comments might not reflect the student's needs. For example, we can imagine a fairly strong essay that garners fairly few Directive comments, many Non-Directive, and no Praise, Criticism, or Off-Topic comments. In such a case it might be tempting to conclude that the instructor has chosen to comment on grammar and has largely ignored content, but it's also possible that the essay was a fairly strong one in terms of ideas, but has numerous grammatical issues. We can easily imagine an ESL student who requires more attention to style and grammar simply due to his/her unfamiliarity with the language.

Limitations to the Study

It might be easier to compile a list of ways that this study wasn't limited, rather than ways in which it was. More than anything else, the difficulty of finding enough
subjects to participate, and having them participate in the ways that a proper protocol would demand, was overwhelming.

Looking back at my design, in some ways I wonder if a textbook protocol was doomed from the beginning. Most studies in academia involve unaware subjects. That is to say, the average freshman pulled from a psychology department's subject pool doesn't have much of an opinion about the studies that he or she is about to take part in. That subject probably has little to no knowledge of psychology, so can't form a reasonable hypothesis on the outcome of the study. One experimental group seems as good as any other.

My subjects, on the other hand, were subjects taking part in a field within their area of expertise. While most of them were not experts in the field per se, they'd all done graduate coursework in composition and were in the act of teaching while they participated in the study. As such, it's difficult to imagine that they did not have some idea of which commenting style would provide benefits above the other. Part of the job of a graduate study, as an authority-in-training, is to have opinions on such matters to drive research and critical inquiry. In this sense, it's no wonder that instructors refused to take part in any random assignment to a research group. To do so might have at least implicitly questioned their authority over the classroom or their knowledge of composition pedagogy.

Yet when I look at the research in hindsight, I'm not convinced that full cooperation would have solved my problems, either. While the protocol itself would have had more integrity, the number of participants still would have been too low to confidently make conclusions. Yet how many instructors would be necessary? On its
simplest level, the study should contain somewhere between 10 and 30 instructors for both electronic and traditional commenting. For simplicity's sake, that would be about 40 instructors. Yet over the course of this study I examined three relatively distinct groups: graduate assistants, full-time MAs, and PhDs. Each of these groups may have had different commenting characteristics worth studying.

At this point the number of participants would triple to 120 instructors. Finding that many participants in an adjunct pool would be difficult enough. Finding 40 PhDs to participate would be nearly impossible. A minimum of 4-5 universities would need to collaborate on the study, and this too would raise the study's complexity.

One of the constant themes throughout this study was the wide range of pedagogical approaches and individual style that composition classes display. Even in this relatively small sampling of Writing I classes, I ran into several points where either the essays or the ways that instructors chose to give feedback didn't fit into my coding rubric. Before beginning this study, I'd never imagined group essays or incorporating graphical images as an integral part of the assignment. This didn't even cover the full extent of possible comments that students could receive. During my time at YSU in the MA program, I've received feedback from instructors in audio MP3 form, and essay feedback as a terminal comment in letter form with no marginal comments. Some classes didn't assign essays, but instead consisted of writing projects such as short stories or group presentations where the only written work was a PowerPoint presentation. I've heard of courses where instructors give no written feedback at all and instead rely on conferences to do the same pedagogical work.
Fortunately, the majority of Writing I courses don't favor such exotic assignments or feedback styles, but nonetheless we must be aware of such possibilities. If the study were to expand, we might begin to see more essays that incorporate visual literacy, audio component, possibly performance and presentation components, or other multi-modal approaches. This seems unlikely, given that Writing I is generally so centered around essays of one sort or another, but if the study has already run into odd cases then it seems that expansion would inevitably exacerbate these issues.

Once multiple campuses enter the research methodology, coding issues become even more apparent. Writing I at Youngstown State University is the introductory writing class meant to prepare students for more advanced forms of academic writing. The primary purpose of the course is to introduce students to academic writing and prepare them for Writing II, the argumentative research-based writing course at Youngstown State. Though many colleges use a similar two-course composition sequence, such congruencies are not guaranteed. The question then becomes what model, and which class in that model, to base the study on.

Recommendations from Phase Two

At the end of Phase Two, coming to conclusions about what this research means is difficult. At first glance, the data don't have many clear trends, and the ones that do exist contradict each other. Which is better: more comments or longer comments? Those answers aren't clear, and unfortunately this study did not attempt to correlate instructional methods with student performance (such as grades or graduation rates).
But even in the face of no clear trends, some pedagogical implications can still be seen in these data. The very lack of clear trends may be a comfort to many instructors. The general trend in the literature has been to emphasize the advantages of electronic commenting. The general feeling that technology will improve classroom practices is not surprising. Though we've all had moments of nostalgia, and many people cling to the notion "paper is just better" or "the smell of books just makes me happy," those claims are accompanied by the understanding that we all have Luddite leanings in a world that's inevitably moving forward. Technology, presumably, is making the world a better place. People are living longer, producing more, and more information is available to the average student now than a hundred years ago. Indeed, even as English instructors swear their undying allegiance to physical books printed on good old-fashioned paper, it's unlikely that any of those instructors still pine for the days hunting through card catalogs. I sincerely doubt that anybody bemoans the advent of searchable databases of academic journals.

Yearning for "the good old days" is impossible to separate from the implicit understanding that technology eventually changes how we undertake nearly any task. From that point of view, it seems inevitable that technology will change how instructors create feedback. The important questions are how and when. Of the two, it's more popular to discuss the "how" but it's just as important to consider the "when."

As the observed data stand, at the moment there's no pressing need for instructors to change their feedback style. If the commentary styles offer mixed benefits, then there's no reason to pressure instructors to switch to electronic commenting. Or equally, there's
no reason why instructors who would like to switch to electronic commenting should be forced to use one style.

Suggestions for Future Research

Given the difficulties observed in this study, one possibility for continuing research into instructor feedback styles is to control the situation even more. This might be possible by creating "dummy" essays that each instructor could provide feedback for. Creating a standard essay that all instructors would examine would control for esoteric assignment types as well as differing course models across different institutions. Presumably, instructors could also be provided with an assignment sheet that described the (hypothetical) context of the class. A similar model was used by Richard Straub in his examination of instructor feedback in *Twelve Readers Reading* (4-6).

Unfortunately, this approach has tradeoffs all its own. The study would need to carefully balance the number of essays that instructors would be asked to comment on. While the study might have good luck finding instructors willing to provide feedback for one essay, such a limited sampling of the instructor's comments might yield dubious results. The previous research has already established that instructors tend to be highly variable in their feedback. With such a high variance, it would be difficult to achieve sufficient statistical power to measure small or subtle differences between traditional and electronic feedback. It would be possible to overcome these difficulties somewhat by creating a battery of 5-10 dummy essays that would provide a more holistic view of each instructor's comments, but as the number of essays to evaluate rises the objects to
participation would also rise. This wasn't as much of an issue in the previous research because instructors were writing comments for essays that their students had written during the normal course of the semester. That is to say, these were essays that the instructors were going to have to write comments for anyway. It was simply the mode (and often not even that) that changed. Instructors who provide feedback for essays purely for a study would see themselves as highly inconvenienced by the study and likely decline to participate.

Additionally, this approach risks biasing the data generally by providing a set basic profile for comments. For example, a dummy essay that instructors perceive as extremely poor may be less prone to elicit Praise comments from instructors. On the surface this might not appear to be a significant problem, given that the study aims to examine comparative differences, but if Praise comments appear unwarranted to too many instructors then the differences between the two commenting styles might appear smaller than they otherwise would. Thus, the dummy essay would need to be sufficiently lengthy and complicated to establish reasonable cause to use each type of comment category.

More than anything else, future research conducted under this model would face potentially unacceptable limitations due to its inability to generate conclusions about how feedback will influence student performance. Though the previous research didn't explore this idea in detail, determining how students react to them is the ultimate goal for any study of instructor feedback. By removing actual student essays from the study, and without an actual student's performance to measure or reflective feedback from the student we would be unable to make any conclusions about which commentary style is
actually more helpful. It might be necessary to proceed with such a simplified study in
order to establish a knowledge base to work from, but nonetheless it would be
disappointing to take such a step.
Chapter Three: Thinking Bigger by Thinking Smaller:
Examining the Approach and Interpreting the Methodological Ideology

In the first and second chapter, I described a research study through its original conception, through the data gathering and expansion, and through the conclusions. Ultimately I've shown how the study generated some useful results by demonstrating that instructors should not feel pressured into using word processors to write feedback. But that having been said, I think the bigger conclusion is that this research into teacher feedback would require much more data to make the kinds of claims that my original study aimed to generate. While a conclusion that more data is necessary is a disappointment to be sure, I'd like to take some time now to reflect on the study and think about how I envisioned this study, what effects that research mindset had, and ultimately how we can re-see this study as generating very valuable results despite what might look like a very limited dataset.

When I took a step back and look at the entire project as a whole, I was not sure what to think. The first thought off the top of my head after running all of the numbers was to declare the entire project the most glorious failure I'd ever had. Although I'm sure my friends, professors, and mother would assure me that my research is a special and unique snowflake, in many ways it was a failure. At the end of the study, the most important conclusion I drew was that instructors shouldn't need to feel pressured to adopt new technology simply for the sake of moving towards what many might assume is the inevitable march of technology. Though educators and administrators might assume that
we have to adopt new technologically-aligned practices and utilize the latest tools, the fact of the matter might be that instructors generally comment along the same lines regardless of whether or not technology assists the comment-writing process.

These weren't bad results per se, and in a sense I'd achieved a measure of success by generating data that argued against the prevailing view that technology inevitably improves practices. But even so, I found the research's conclusion a bit disappointing.

Whether or not anybody likes to admit it, we all know what the sexy research is. It's posing a question that nobody ever thought to ask, and it's researching a phenomenon that suggests how "this writing thing" ought to get taught. We all know who gets the cover page on the composition calendars. That privilege belongs to researchers whose findings suggest new insights into pedagogy. Deep down, I suspect every researcher would like to create a new foundation for a different theoretical model that sweeps the nation (and journals). That's the goal of publication: to say something so important that everybody has to read it.

When I had set out to begin this study, I had some of those sexy results in mind. Those thoughts might not have taken center stage, but I can see them when I look back at Chapter One. On page 3, I wrote, "While these reasons may resonate with many instructors, the most likely reason why instructors have not switched over to electronic commenting is institutional inertia. Quite simply, the system is not obviously broken, so there's no need to change it." On page 5, I wrote, "But even if electronic methods of commentary are not the norm now, it is important that we begin examining the differences between paper and electronic comments to determine if one form favors the creation of comments that instructors or students prefer." The mindset here is forward-
looking and critically evaluative. I was looking for institutional relevance, and I was looking for the differences in commenting that would indicate the better commenting style.

I can see now, with the clarity of hindsight, that this study's early evolution was dominated and crucially shaped by my early experiences learning research methodologies as a undergraduate in psychology, a discipline traditionally rooted in quantitative research (Gelo et al. 266-267). Beginning my research was relatively easy. I created a fundamental research question, and then found the most logical way to find answers for that question. If I wanted to know which commentary style created more comments ("more" presumably being better), then I'd simply count them. If I wanted to know if the comments were more helpful, then I'd code them according to a rubric designed to categorize comments into distinct categories associated with helpfulness. I identified key points to measure, found ways to control for confounding factors, and then began collecting data.

The problem was, I had no idea how complicated such a study would become, or how complicated an ideal version of the study would be. In my various ruminations about ways to "fix" this study, all of my ideas for overcoming the methodology's drawbacks involved adding yet more participants, collecting more essays, and developing yet more complicated coding schemes. Until now I've only considered the practical reality of finding so many participants, but in truth the drawbacks to an expanded study with many more research participants go much farther than simply finding more participants.

To accomplish even a modest research project such as this one required considerable time and effort. Alice and I began this project as part of a research
methodologies course during the Spring semester of 2009. We further collected and
coded data in Summer 2009, and we both began the Fall 2009 semester engaged in much
the same work (though Alice later departed the project to pursue her own interests with
the data we'd already collected). In total, the project required significant time and effort
for an entire calendar year.

Looking at the kind of work generated by expanding the data collection shows
some of the other concerns that only became apparent later on in the study. Much of the
time required to conduct the previous research involved waiting for essays to be
collected. Assuming highly-cooperative participants and institutional support, that time
frame could be reduced considerably. However, coding the data is a very labor-intensive
process that requires significant training and judgment. The expansions to the project that
I've envisioned over the course of this study would approximately triple the amount of
data, which would require considerable manpower.

Even worse, even with all the measures I've imagined, it's not clear what answers
the project would generate. With 60 instructors commenting on dummy essays, we might
be able to make reasonable claims about electronic versus traditional commenting, but
those comments would have been generated in a very controlled environment that might
have some validity, but probably wouldn't satisfy all researchers interested in the topic.
At some point the research needs to indicate real-world situations, or at least mimic real-
world conditions closely enough to indicate new best-practice teaching strategies.

On one hand, all of this is possible. It's perfectly possible to use the research I've
imagined as a beginning step towards expanding research to eventually ask those
extremely difficult questions. In many ways, this is simply the pattern that researchers
use to pursue areas of inquiry. Small studies lead the way to bigger studies, and often those initial investigations include control factors or simplifications that reduce real-world applicability. There's a common joke in the sciences that goes something like the following:

A farmer sees that his cows aren't producing enough milk, so the farm isn't making enough money to pay the bills. He goes to a physicist and asks to help. The physicist goes off to his lab and works furiously for a week. He then calls the farmer and explains that he has the answer. The farmer goes to the physicist’s lab for a presentation, and the physicist begins by saying, "First we consider a spherical cow of uniform density…”

As Chad Orzel notes, this joke is only really funny to physicists, but it does do a good job of poking fun at the way that research examines carefully-crafted, patently ridiculous notions in order to understand basic principles before adding the complexities of real life back into the situation (Orzel). Even simple research studies can pave the way for more complicated studies, and as I moved through Phase One and Phase Two of this research this was the mindset that motivated me. At every step I asked myself what the next step—the next expansion—would be. Even at the end of Chapter Two, when I imagined an even more complicated Phase Three, I still wasn’t entirely satisfied with the methodology that I’d created. There were vast issues that I hadn’t yet considered in depth, such as more advanced and detailed ways to code comments. At the very least, breaking the grammar marks into categories seemed necessary, or at least helpful. But as I added each step of reality back into the study, the enormity of the task became ever-more apparent, and the less results the study was actually capable of generating.
The way that I developed this methodology can be traced back to two
fundamental causes: First, my own inexperience with conducting research. Second, the
quantitative ideology of research that I'd been exposed to as an undergraduate studying
psychology.

When I look at the early chapters of this study, I'm somewhat surprised by what I see. The early genesis of the methodology was straightforward, even easy. That ease may have actually been the first sign that the methodology had subtle flaws. Much in the same ways that writing is easy when authors slip into the pre-formed thoughts encapsulated by clichés, my study also fell into well-worn patterns that made it easy to design the study. The only difference was, my clichés came from the positivist/constructivist ideology associated with quantitative research. From the earliest days, I was as interested in asking as "big" a question as I could: I wanted to know how writing instructors (as though such a wide group could be boiled down to a single essence) should teach. I may have described the study as an examination of traditional and electronic comments, but that was only a fairly benign way of describing my purposes. My real purpose in studying electronically-generated comments was to see which one was "better."

At the time, I hadn't rigorously defined exactly sure what "better" meant, though the metrics I was looking at indicate my priorities. The easily-quantifiable dimensions received the most attention at first. It was quick and easy to count the word-length of the comments in the essays I examined. It was equally easy to count the number of comments. Even better, these qualities were associated with helpfulness for the student. After all, more instances of advice would naturally be more helpful than fewer. The
longer an explanation, presumably the fuller and easier to understand. A longer comment meant more detail, more nuance, and possibly more advice for how to revise.

But that's not necessarily true, and even a few moments of reflection would result in a critical reader finding problems. A longer comment might be more helpful, but it might also be a sign that the instructor rambles. Maybe the instructor is simply wordy, and the additional verbiage actually confuses the student. Or perhaps a large number of comments per page is actually overwhelming. Perhaps students automatically assume that a lot of comments means there's too much work to do with an essay. Such a student would be inclined to scrap the essay and start over from scratch rather than revise, or possibly might stay with the current draft and assume that the work required to receive a better grade wouldn't be an efficient use of time that could be equally used for other classes or activities.

"Better" has to be defined in terms of the study's goals. Better for whom, I might now ask my prior self. When I look back at what I wrote in Chapter One and Chapter Two, I'm not sure if the term "better" was well-defined. Does "better" in the context of this study mean that the instructor is able to better explain the effective and ineffective parts of the essay? It might mean that the instructor is able to express the requirements or expectations of the assignment. Alternately, "better" comments might express the instructor's visceral feelings of confusion or comprehension.

Another possibility entirely is that "better" actually refers to helpfulness as perceived by the student. "Better" comments might refer to legibility. It might also mean that the student is encouraged to revise the essay. It might mean that revisions are more effective, and result in an essay that is more likely to educate or persuade a reader. Or
effective comments might mean that the student is able to conceptualize some aspect of writing to recognize the elements that prompt the instructor to write comments, thereby helping the student to write more effectively on a future essay.

All of these definitions aim at what I implicitly had in mind when I created this study, but looking back I'm not sure exactly how those nascent thoughts interacted with each other. Though all of the ideas I just mentioned relate to each other, they don't always occur together. An instructor who communicates his feelings about the writing might not encourage the student to revise, and a student who revises well for one essay may have trouble transferring those skills to a different piece of writing. With that in mind, I'm not sure if a goal of determining the "better" method of writing comments was ever really achievable, if for no other reason that because I hadn't properly defined that term and thus didn't have the kind of instruments or measurements that would have led to such an answer. To embark on that study would have required many smaller studies working in tandem to build up to increasingly complex ideas of what writing is and describing the various interactions that take place between a teacher and student, student and the essay, and the teacher-student-essay relationship in the context of a classroom full of other students with simultaneous relationships of their own.

When I re-read my first two chapters, I see confusion where I originally saw perfect sense. My current self feels like one of the instructors in my study, writing notes in the margins that say, "Are you sure these terms mean what you think they do?" and "I think you're on to a good idea here, but develop this idea. Where does it lead you?" This is not to say that I was entirely insensitive to these concerns when I began this research. Indeed, at the time I thought that I was conducting research to the best of my ability,
which simply included using imperfect solutions to solve very difficult problems. I never stopped to think about the nature of the questions that I asked. I never stopped to think about whether or not my questions were addressable in the time or data set that I had available to me.

I can now see in retrospect that achieving all of the stated goals of the original study would have required much more time and effort than I could commit to the project. The data I gathered simply weren't sufficient to describe all writing instructors. A study can only make claims for as large a pool as it studies. If I were to make claims about writing instructors as a whole, then I needed a participant pool that reflected a representative cross-section of all writing instructors. That is a very difficult standard to achieve, as my search eventually found. Even so, it's not clear if the basic methodology that I used to conduct Phase One and Phase Two was insufficient to the task of describing instructor feedback. I would argue that the concepts of "sufficiency" or "appropriateness" in research methodology aren't very easy to define or apply. My study may have been insufficient because the amount of data I gathered was too small to make the claims I wanted to, but that is an argument against my implementation rather than the fundamental methodology. Quantitative research has the potential to bring important insights to composition. This trend goes back to 1961 when ETS developed the first rubric for writing assessment through factor analysis (Broad 5-9). In that case, ETS attempted to reduce all of writing assessment into five factors. These factors could then be given a convenient numerical rating that corresponded to a Likert scale. These scales are extremely popular in writing assessment, as evidenced by a study of inter-institutional
college writing assessment conducted by Pagano et al. which measured writing on a 6-category scale with scores falling on a 5-point Likert scale (Pagano et al. 294-299).

In many ways, the study that Pagano et al. describe look somewhat similar to the expanded study that I imagined. Participating institutions created prompts for use in the courses that would be assessed. Those prompts were created agreed-upon parameters by committee, thus reducing the context of the courses somewhat (290-292), and the study similarly produced a 5-point Likert scale as the final assessment of writing (292-302).

This kind of assessment may produce answers that answer the ostensible question asked (which school tends to produce the best writers), but it's not clear what else those data can tell us. The assessment in question was created with the best of all possible intentions, given the evidence. Pagano et al. created a scoring rubric based on holistic criteria that reflected the group's criteria of good writing. It was well-controlled, replicable, and expandable. Perhaps more telling are the conclusions that these data do not tell us. A score of 3.56 on a given metric is certainly superior to a 3.36, and that difference can be shown to be statistically significant, but that doesn't tell us much about the writing. Even if the scoring rubric was generated according to holistic means, where scores of 1 were equated with characteristics such as "Thesis statement is obviously logically flawed or non-existent," that information is lost in the final analysis. Indeed, it's not clear what the resultant data mean. While the raw data can be traced back to the coding scheme easily, it's not clear what an aggregated mean of 3.36 tells a researcher. Such a score is worse than a 3.56, and again still worse than a 3.76. But these scores are not immediately useful for a teacher. They tell us nothing about the individual essays.
Indeed, it's not clear that these scores have any meaning outside of a mathematical description of how data interact or how means are generated.

Likert scales are helpful because the data are easy to gather through surveys or coding rubrics, are easily manipulated through statistics, and are quickly and easily reportable, but it's exactly those same qualities that make them unhelpful for extracting meaning out of those scales. The Likert scale is what engineers might call a "lossy" transcoding. That is to say, not all of the information that went into the final analysis can be extracted from the analysis. Information is lost in the very nature of the data analysis. We can't even reconstruct a picture of a typical student essay by analyzing the Likert scale because differences may be measured in the tenths or hundreds of a point. Imagining this as a prototypical essay would mean changing a few words, perhaps a comma. Translating the score into a meaningful real-world terms would be impossible except in a case where the difference in score were fairly large, perhaps on the order of 10% of a scale's range. So while the Likert scale might be the ideal tool for measuring small differences in broad trends, it's not even clear what that small difference means.

While that loss of fine meaning is a bit unsettling, for the most part I find myself untroubled by it because I find the tool useful for what it is: for broad evaluations of highly complicated phenomena. A Likert scale is a good way to describe trends in very complicated situations. To ask a Likert scale to make fine distinctions simply asks the scale to do a job that it was never designed to do, and I can accept that. As I mentioned earlier, my research paradigm was primarily informed by an undergraduate degree in psychology, a behavioral science firmly rooted in quantitative research. That training told me that people are too complicated to assess as deep, detailed individuals. To that end,
It's helpful to create coding rubrics that simplify complex interactions and systems into a form that can be more easily examine. That is to say, turning a cow into a spherical cow of uniform density, or in this case turning a large set of rich, detailed, complicated essays into a single number. But it would be a mistake to dismiss all quantitative research on this basis, because Pagano's study isn't meant to draw conclusions about individual students. Nor is it really designed to discuss the improvement of an individual student's writing. Indeed, it's not even clear that the study is designed to indicate how any given institution should improve student writing. The study specifically offers a method to compare one school to another, and insofar as it accomplishes this relatively small task, it does so well.

The question here isn't so much whether or not the methodology that I selected was correct. What's more important is the kinds of questions that I was asking, and whether or not I was examining the data properly. Though the questions I was asking were complicated, and require a great deal of judgment, that does not necessarily mean that I should have pursued the research in a fundamentally different manner, such as a qualitative study. As Michael Westerman notes, all quantitative data is inherently subject to interpretation. For example, the coding systems that quantitative researchers use assume a wealth of information in order to make accurate and trustworthy decisions about how an infinite range of possibilities can be divided into discrete categories (191-193). There's nothing inherent in the asking of what "better" comments are that would have precluded searching out these answers, and indeed I can imagine designing many quantitative methodologies that would have addressed different aspects of the question. For example, survey data could provide information about what kind of feedback students would prefer to receive. Student performance could be tracked by examining
final draft grades, then correlating higher grades with a feedback style. As a follow-up, students could be tracked longitudinally to determine if receiving electronic or traditional feedback in a Writing I course has an effect on success in Writing II, a writing-intensive course in a Writing-Across-the-Curriculum setting, or even final graduation.

The problem that my study faced was that it started with a question that was too broad to adequately address with the available manpower, and in a reasonable amount of time. In retrospect this may not be surprising, given that quantitative research is somewhat more inclined to attempt to answer research questions of very ambitious scope. As Omar Gelo et al. note, quantitative research has long been understood to consist of "the establishment, collection and assimilation of facts with the exclusive aim of recognizing and formulating laws that are always and in every circumstance immutable and universally applicable…" (270). In other words, a research question such as "How should all writing instructors write their comments?" is entirely possible (even favored) under a classic view of quantitative research because the research has historically sprung from an ideology that states that such broad proclamations are the highest goal.

Here I'd like to specifically note that I am not arguing that quantitative research was wrong or inappropriate for this study. At its core, quantitative research is a set of research tools that rely on quantification of phenomena and the statistical methods of analyzing those data. As such, there was nothing inherently wrong with the methods. Rather, the problems that I experienced came from the ideology of research associated with the methodologies, and the perception of the quantitative-qualitative debate is largely the result of disagreements between the ideologies that form the use of those research methodologies (Gelo et al. 268-270). This re-envisioning of how quantitative
research techniques can be used also closely relates to Linda Adler-Kassner and Heidi Estrem's notion of changing the frame, where they argue that ideas of a common notion such a student writing carries exist within a "frame" that encapsulates implicit notions of what student writing is and what it is capable (or incapable) of, and therefore what most composition instructors think of it (19). Frames are useful concepts because they give people a quick, easy way to think about concepts. The frame is a metaphor that carries values and assumptions about how parties interact with each other. They explain what actions are appropriate, as well as which ones are not (Lakoff 7-14).

The notion of framing is most useful for it being a frame itself. That is to say, we can imagine an activity such as research as taking place within a conceptual movie frame. The conceptual center of the frame defines what we focus on (that which is valued or seen as culturally appropriate), that which is near the edge is not valued or not culturally appropriate, and there are even out-of-frame concepts that are uncommonly thought of. These are the so-to-speak "out of the box" thoughts. These frames work very well for ideas that translate into metaphor well, such as Lakoff's example of politicians invoking the image of a permission slip when describing the United States' decision not to consult the United Nations before invading Iraq (11). Such models draw on the implicit knowledge gained by membership in a culture and make discussions faster and easier. Research doesn't have such an easy metaphor. Without this easy reference, frames become merely another way of talking about ideologies. The most important point to keep in mind is that ideologies suggest methods of action as well as values. Once this aspect is emphasized, frames and ideologies express the same basic idea, and as much as
I like the idea of changing the frame that Estrem and Adler-Kassner use, I prefer to consider this issue in terms of ideologies and paradigms.

I can see some of the way that my research mindset changed over the course of the year and a half since its beginning, and integrating the two research styles begins to show that change in my thinking. In Phase Two of the research, I did collect a mass of additional data--interviews and surveys of instructor and student satisfaction with the comments--meant to triangulate my results and start to answer some of the questions that might have led to the kinds of answers that I originally asked for. A very similar method has been advocated by Gelo et al., who argue in favor of "mixed methods" research drawn from the worldviews of both quantitative and qualitative research, where quantitative and qualitative methods of data gathering are combined in a single study either in parallel or sequentially to provide a greater understanding of the subject (278-286).

The danger inherent in viewing this divide as a debate is the risk of falling into a war of ideologies that dichotomizes research when instead the focus should be on finding the right ways to investigate our research questions, and even beyond that ensuring that we ask the right questions. Fortunately for me, my study unintentionally included some of the checkpoint milestones that I described earlier in this study. At the end of Phase One, my study had generated good, clear results that indicated trends in the populations I had examined. Continuing by expanding the study to address the methodological simplifications I'd made to control the data was a logical next step. Phase Two ended after a limited data collection that showed just how much more work I would have to do in order to achieve my original research goals, and as difficult as that was to accept, it
also provided me with a valuable opportunity to look at the data once again and try to see the study in a different way. It allowed me to ask what I had already done, rather than simply what I still wanted to do, and what those data actually told me rather than what I wanted them to tell me.

When I looked at the study through this lens, I found that I hadn't found the results that I had been looking for, but I did find something potentially even more valuable: I found what each instructor had experienced over the duration of the study. In the case of my four participants who provided comments in both styles, I had case studies showing how each instructor's comments changed depending on the feedback modality. This extremely small context offered many advantages that weren't obvious at the beginning of the study, and had in fact appeared to be disadvantages when I began. When I started this study, I assumed that looking at one instructor would be useless. I saw a single instructor as a single data point. But by reinterpreting the results and re-imagining the research questions, I saw each instructor as a separate study. This changed the meaning of the data and showed how the limited scope controlled for many factors. The results, by their very nature, took into account most contextual factors that would have founded a more comprehensive study. Instructors tended to assign essays of comparable length and taught in a consistent context. Leon and Fran submitted rough drafts for both phases of the research, so they were able to control for variances in the amount of feedback instructors give rough drafts vs final drafts.

These individual instructor profiles (summarized in Appendix A-D) illustrate a more balanced research mindset: Rather than follow a prescriptive notion about what questions the research should ask, the methodology instead attempts to describe what the
available data are capable of saying. Instead of asking what's the best way to discover the answers, the emphasis becomes looking at what the research is capable of generating and asking what answers can be found. Shifting focus allows the research to avoid tunnel-vision and instead places focus on keeping perspective.

At first glance it might seem that this approach lacks the power and potential of traditional research methodologies. The context of these new findings was certainly smaller in scope than what I originally set out to prove. By limiting the context of the study to single instructors, the question "What's the best/most helpful feedback style?" ceases to make sense. The question can't be addressed due to an overwhelming lack of data. But it does answer a different question: "What happens when this instructor switches from writing traditional comments and goes to electronic comments?"

Discovering the most effective feedback style for a single instructor may not seem important, but we must remember that this isn't precisely a single instructor; it's any given instructor. That distinction is important because instructors can use this research to guide how they offer student feedback. The individual instructor profile can tell that instructor that he/she tends to write more Directive comments and fewer Praise comments when using electronic means. Or an instructor might tend to make more grammar marks and fewer Non-Directive comments when hand-writing feedback.

This level of detail is important because it allows instructors to choose the most effective commentary style based on the student's needs and the assignment. For example, an instructor might want to focus on Directive comments for rough drafts where revision is extremely important, but then switch for the final draft where emphasizing Praise comments is more helpful. Instructors might also want to switch styles for a
particular student. For example, strong students whose grammar errors reflect a lack of proofreading rather than misunderstanding grammatical concepts might be better served with traditional comments that identify errors, while an ESL student who needs to be taught those rules would be better served with a fuller Non-Directive comment. In this sense, the two styles do not reflect a "good/bad" or "better/worse" dichotomy so much as an appropriateness for the stage of writing and the student's individual needs.

Even better, this same effect can be replicated for other instructors. Hypothetically, an instructor who teaches two sections of a class could track his/her own feedback, coding it according to the rubric, and give feedback in one style for one section, then provide feedback in the other style for the second section. Coding approximately 40 essays in this manner would require around six hours, assuming an instructor codes an essay in about 10 minutes; this does not include the time to write the initial comments. Instructors could even be provided with a spreadsheet set up with all of the relevant statistical calculations, which would make analyzing the data effectively instantaneous.

For the investment of six hours, an instructor can get an analysis of his/her comments in two styles. Even better, that work can be completed over a long period of time. The work could be completed over one session of six hours just as easily as three sessions of two hours. Completing the analysis over two weeks, with a half-hour each day, is also possible. The rubric doesn't require much in the way of specialized tools, so it's fairly cheap to implement.

Most important, this method still leads to the one answer I most wanted at the beginning of the study: What should instructors do to best teach their students? Originally
I thought that the best way to get these answers was to research the feedback styles and make a recommendation that would suggest how teachers should be trained. I was looking for strong results that would inform departmental policies and suggest new ways to imagine the composition classroom. But in some ways these results are superior to broad, bland policy recommendations. The instructor profiles offer much richer information than simply saying "Instructors tend to generate 50% more comments, which are 50% longer." Each profile contains a rough description of how the comments change.

It's important to note here that even the revised approach to research that I've described still starts in much the same way as this study originally did. That is to say, the research still begins by formulating research questions and attempting to find the answers to those questions. The main difference is that at some point the research stops to evaluate its progress. I see now that research projects should incorporate regular "checkpoint milestones" where the research can stop and researchers can examine the data and asks if there are already useful findings in what's been gathered so far. At this point the researchers can assess the methods already used and determine if they're sufficient, insufficient, or simply wrong for the questions already asked.

Stopping to asking "what do we have?" or "is there another way to ask this question that might be supported by the data we already have" is the most visible benefit of this approach, but there are other pragmatic benefits as well. If the research begins with a small-scale study, initial problems can be detected before too much time or too many resources have been spent in data collection and analysis. Even this study shows wide data-gathering can fail to be as helpful as expected. Though I collected interviews and surveys from instructors and students alike, none of those data were eventually
analyzed for this study. The analysis itself didn't make much sense because the results
from Phase Two were initially inconclusive. As a result, the surveys and interviews
weren't able to place the (non-existent) results into context. Fortunately, in this case I
didn't devote considerable time or energy interpreting those results. That time and energy
were better spent on other aspects of the research. Assessing progress at small junctures
eventually creates a more efficient deployment of resources, at the possible expense of
time.

In many ways, this iterative approach looks similar to qualitative approaches to
research, such as the ethnographies described by Wendy Bishop, wherein researchers
conduct data collection with only preliminary research questions or a nebulous sense of
the desired findings (41-46). Ultimately, this points to the convergence of qualitative and
quantitative research that is not so much indicative of the methods themselves but the
ideologies that created the two methods. Paying attention to these ideologies is crucial to
a project's long-term success because ideology affects a study's design as much as any
data collection method or statistical tool. I can see now that many of the choice I made
were simply reflex decisions made as a result of the ideology that I was trained in while
learning laboratory research methods as an undergraduate (a tradition that heavily
emphasized quantitative research and an ideology that valued asking Big Questions and
generating Big Answers above all).

I have to admit, I got caught up in over-engineering the original research. I
thought that by doing more work at the beginning, I could find results that would have
fast, easy implementations. What I found was a very different story. I have a story that
says that some of this research will be ongoing, and that instructors who are willing to
participate will reap the rewards. Perhaps not surprisingly, the instructors most willing to reflect on their feedback and analyze their practices have the most to gain. The instructors willing to analyze their comments from similar assignments, in similar situations, across multiple sections, can generate very good data that lead to non-confounded results.

But even though instructors may create individual feedback profiles that appear to benefit only themselves, this would not stop the research from moving forward. The system that I've imagined is modular and scalable through the accumulation of data over a long period of time. Were we to finally collect data from hundreds of instructors, the kinds of research questions that I originally tried to answer might still be answerable. While one data point is merely anecdotal, hundreds or thousands of data points could form the basis of a large-scale meta-study that could never be achieved by the kind of command-model usually envisioned by researchers. This meta-study could serve as the basis for a broad baseline description of what instructor feedback looks like across multiple institutions, taking many different contextual situations into account. With that baseline in place, we might begin to imagine new ways to describe the data we have to determine if expansions to investigate interactions with student grades and other measures of performance are feasible. Such a study would be impossibly difficult to organize from the top. Instead, we must rely on instructors to do some of the work on their own. In most cases that approach wouldn't be sound because of the low payoff for the work, but in this particular case the model offers instructors an immediate personal benefit.

Of course, instructors may not wish to replicate this research on their own. Though the research would only require six hours, instructors would need to attempt the
new commenting method (or revert to an old one) in good faith. That will naturally require a certain amount of risk, and most composition instructors already struggle with heavy workloads that offer little time for reflection and research. Given the choice between research and allotting the same time to proven practice, most instructors will choose the proven practice. The choice is understandable. Most instructors are well-familiar with their commenting style. They know how long it will take to write comments for each paper, so that risk is low. Even if the reward for switching to a different commenting style is present, the magnitude of the reward is uncertain. The six hours of research time would only be justified if it led to productivity increases that offset those six hours in a reasonable time, and doubtless some instructors would want a very fast return on their investment of time. Such results are unlikely, so the long-term benefits of research are made secondary to the ever-pressing deadlines for grading and giving feedback. Long-term strategy is sacrificed for short-term tactical needs, as a military analyst might say.

I don't think that there is a perfect way to increase participation in research. Certainly administrators could offer faculty members release time for participating in research, although I suspect that such measures would have a limited effect. Indeed, most of the usual institutional encouragements for research don't reach the most-needed population: the adjunct professors. Though these instructors teach the bulk of composition sections at many universities, there is very little incentive at the institutional level to participate in research. Here, the immediate personal benefits to the instructor become the best argument in favor of participation, but I fear even that may not be
enough. Adjuncts exist at the fringe of academia, teaching in its classes but for the most part not participating in its committees or research.

Ultimately, if what I've suggested in this chapter is that we need to adopt new mindsets when conducting research, then some of those changes may need to become more systemic through the university. Seeing research as a collective responsibility may well be one of the steps that we need to take. If we all stand to benefit from the research, then it is only fitting that we should bear shared responsibility for that research as well. Undertaking a comprehensive project, across multiple institutions and involving hundreds of instructors, would be far too difficult for a single person or even a small team to undertake. In such a situation, the best that I can do is to organize future research. I hope that the individual benefits available to instructors who choose to participate in the research are enough to provide the initial impetus. But it may not be enough, and if that turns out to be the case then it will be a sad conclusion for us all. It will mean that to some degree, we've all failed each other.
Works Cited


Orzel, Chad. "Assume a Spherical Cow."
http://www.steelypips.org/principles/2003_03_30_principlearchive.php#2000983


Appendix A: Individual Instructor for Bill

<table>
<thead>
<tr>
<th></th>
<th>Electronic</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td>* significant at p&lt;0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>** significant at p&lt;0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean comment length**</td>
<td>15.00</td>
<td>6.83</td>
</tr>
<tr>
<td>Mean comments/page</td>
<td>3.42</td>
<td>3.74</td>
</tr>
<tr>
<td>Mean grammar/page**</td>
<td>1.53</td>
<td>0.16</td>
</tr>
<tr>
<td>Mean terminal comment**</td>
<td>192.38</td>
<td>252.05</td>
</tr>
<tr>
<td>Mean % Directive</td>
<td>55.22</td>
<td>55.73</td>
</tr>
<tr>
<td>Mean % Non-Directive</td>
<td>25.30</td>
<td>26.97</td>
</tr>
<tr>
<td>Mean % Praise**</td>
<td>3.30</td>
<td>13.38</td>
</tr>
<tr>
<td>Mean % Criticism**</td>
<td>13.24</td>
<td>3.21</td>
</tr>
<tr>
<td>Mean % Off-Topic</td>
<td>2.94</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Appendix B: Individual Instructor for Fran

<table>
<thead>
<tr>
<th></th>
<th>Electronic</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td>* significant at p&lt;0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>** significant at p&lt;0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean comment length**</td>
<td>5.98</td>
<td>5.68</td>
</tr>
<tr>
<td>Mean comments/page</td>
<td>2.28</td>
<td>3.57</td>
</tr>
<tr>
<td>Mean grammar/page**</td>
<td>0</td>
<td>2.14</td>
</tr>
<tr>
<td>Mean terminal comment**</td>
<td>60.79</td>
<td>116.43</td>
</tr>
<tr>
<td>Mean % Directive</td>
<td>38.64</td>
<td>26.55</td>
</tr>
<tr>
<td>Mean % Non-Directive</td>
<td>29.43</td>
<td>46.66</td>
</tr>
<tr>
<td>Mean % Praise</td>
<td>19.12</td>
<td>8.98</td>
</tr>
<tr>
<td>Mean % Criticism</td>
<td>9.70</td>
<td>14.96</td>
</tr>
<tr>
<td>Mean % Off-Topic</td>
<td>3.12</td>
<td>2.86</td>
</tr>
</tbody>
</table>
### Appendix C: Individual Instructor for Leon

* *significant at p<0.05  
** significant at p<0.02

<table>
<thead>
<tr>
<th></th>
<th>Electronic</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean comment length**</td>
<td>16.23</td>
<td>9.15</td>
</tr>
<tr>
<td>Mean comments/page</td>
<td>3.18</td>
<td>4.02</td>
</tr>
<tr>
<td>Mean grammar/page*</td>
<td>0.89</td>
<td>3.31</td>
</tr>
<tr>
<td>Mean terminal comment*</td>
<td>209.63</td>
<td>120.7</td>
</tr>
<tr>
<td>Mean % Directive**</td>
<td>39.67</td>
<td>59.54</td>
</tr>
<tr>
<td>Mean % Non-Directive**</td>
<td>52.37</td>
<td>26.20</td>
</tr>
<tr>
<td>Mean % Praise</td>
<td>5.41</td>
<td>6.92</td>
</tr>
<tr>
<td>Mean % Criticism**</td>
<td>0.96</td>
<td>1.10</td>
</tr>
<tr>
<td>Mean % Off-Topic</td>
<td>1.59</td>
<td>5.04</td>
</tr>
</tbody>
</table>
Appendix D: Individual Instructor for Nancy

<table>
<thead>
<tr>
<th>* significant at p&lt;0.05</th>
<th>** significant at p&lt;0.02</th>
<th>Electronic</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean comment length</td>
<td></td>
<td>11.07</td>
<td>10.53</td>
</tr>
<tr>
<td>Mean comments/page**</td>
<td></td>
<td>1.59</td>
<td>2.71</td>
</tr>
<tr>
<td>Mean grammar/page</td>
<td></td>
<td>1.72</td>
<td>0.73</td>
</tr>
<tr>
<td>Mean terminal comment**</td>
<td></td>
<td>7.23</td>
<td>29.64</td>
</tr>
<tr>
<td>Mean % Directive**</td>
<td></td>
<td>56.55</td>
<td>75.25</td>
</tr>
<tr>
<td>Mean % Non-Directive</td>
<td></td>
<td>17.31</td>
<td>7.50</td>
</tr>
<tr>
<td>Mean % Praise</td>
<td></td>
<td>10.74</td>
<td>14.23</td>
</tr>
<tr>
<td>Mean % Criticism</td>
<td></td>
<td>0.55</td>
<td>1.51</td>
</tr>
<tr>
<td>Mean % Off-Topic</td>
<td></td>
<td>7.16</td>
<td>3.02</td>
</tr>
</tbody>
</table>
August 31, 2009

Dr. Kevin Ball, Principal Investigator  
Ms. Way Jeng, Co-investigator  
Ms. Jolene Stieb, Co-investigator  
Department of English  
UNIVERSITY  

RE: HSRC Protocol Number: 13-2010  
Title: Annotations on Student Texts: An Analysis of Traditional and Electronic Comments  

Dear Dr. Ball and Ms. Jeng and Stieb:

The Human Subjects Research Committee has reviewed the abovementioned protocol and determined that it is exempt from full committee review based on a DHHS Category 1 exemption.

Any changes in your research activity should be promptly reported to the Human Subjects Research Committee and may not be initiated without HSRC approval except where necessary to eliminate hazard to human subjects. Any unanticipated problems involving risks to subjects should also be promptly reported to the Human Subjects Research Committee.

The HSRC would like to extend its best wishes to you in the conduct of this study.

Sincerely,

Peter J. Kasvinsky  
Dean, School of Graduate Studies  
Research Compliance Officer

PJK/cc  
c: Dr. Gary Salvner, Chair  
Department of English