THE EFFECTS OF A SIMULATED SELF-EVALUATIVE ROUTINE ON TEACHERS' GRADES, INTRACLASS CORRELATIONS, AND FEEDBACK CHARACTERISTICS

BY

Charles Hurst Golden

Submitted to the graduate degree program in the Department of Curriculum and Teaching and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Heidi L. Hallman, Chairperson

Phil McKnight

Bruce Frey

Marc Mahlios

Donita Massengill-Shaw

Date Defended: December 8, 2011

The Dissertation Committee for Charles Hurst Golden certifies that this is the approved version of the following dissertation:

THE EFFECTS OF A SIMULATED SELF-EVALUATIVE ROUTINE ON TEACHERS' GRADES, INTRACLASS CORRELATIONS, AND FEEDBACK CHARACTERISTICS

Heidi L. Hallman, Chairperson

Phil McKnight

Bruce Frey

Marc Mahlios

Donita Massengill-Shaw

Date Approved: December 8, 2011

Abstract

English language arts teachers committed to the teaching of writing must allocate substantial time and energy to the evaluation of student essays. And in doing so, these teachers wrestle with at least two star-crossed expectations. First, they must fulfill the institutional obligation of making reliable holistic judgments of the papers they receive, stratifying papers according to their successes against a set of stipulated criteria. Second—and more importantly for the sake of teaching and learning—they must also be the providers of insightful, inviting feedback that promotes rather than hinders students' progress toward robust literacies. The qualities of such feedback, having been studied by Kluger and DeNisi (1996), Hattie and Timperley (2007), and others, have recently been made available to classroom practitioners in Brookhart's *How to Give Effective Feedback to Your Students* (2008). The current study leverages Brookhart's transmission of previous research to investigate how teachers might improve their feedback characteristics by way of a self-evaluation routine administered to students prior to the submission of so-called final-draft essays. Specifically, the study tested teachers' scoring and feedback practices, with respect to their work on stronger and weaker essays across control and experimental conditions pertaining to the absence or presence of simulated self-evaluative comments by student authors. Scoring practices were considered by way of group means, distributions, and intraclass correlations of participating teachers' evaluative scores; similarly these teachers' feedback was coded according to criteria suggested by Brookhart, and then compared by way of a 2x2 ANOVA comparison of feedback variances across stronger and weaker papers under control and experimental conditions. The analyses of these data demonstrated a medium-sized positive effect for the desirable feedback trait of *focus on self-regulation* (partial $\eta^2 = 0.079$), as well as a small-sized positive effect for the desirable

trait of *comparisons to an imaginable previous or successive draft* (partial $\eta^2 = 0.032$). These desirable improvements in feedback were accompanied while maintaining comparative stability in the grades imposed by teachers, limiting the concern that a "friendlier" approach derived from principles in interpersonal psychology (Heider, 1958) might somehow weaken the integrity of rigor in scoring.

For Valerie, whose sacrifices were greater.

For William and Andrew, my best teachers.

And for Leverett, who dreamed.

Acknowledgements

Truly exceptional teachers are always visionaries, having been wrestled down by an angel bearing kaleidoscopic images of a tomorrow-world far more *human* than the worlds into which they have been born. And having imagined this world, their faces shine with a different light than do the faces of their colleagues. Shady Brook Elementary fourth-grade teacher Marilyn Wilkinson was one of these beaming-faced leaders. So, too, were Bedford Junior High School's Avis Neeper and Mike Wharton. At Lawrence D. Bell High School, even those who didn't take his classes *knew* that Bob Stapleton was as good as it gets.

Over the last twenty years, it has been my great joy to have studied under the guidance of several exceptional teachers, including James Barcus, Robert Baird, William Cooper, Jay Losey, D. Thomas Hanks, Jr., David Townsend, Roberta Frank, George Rigg, Phil McKnight, Billy Skorupski, Bruce Frey, Jennifer Ng, and Heidi Hallman. In the workplace, too, I have been graced to know a few true masters. Amy Murphy, Elizabeth O'Brien, Keri Schumacher, John Selzer, and Bill Smithyman are among those who provide constant reminders that I must aim higher and be more generous with my talents if ever my students are to benefit from the sort of instructional leadership that we would all want our children to receive from their teachers.

But those who have had the greatest educational effect on my life are not my "at-school" teachers so much as my parents, my wife, and my children. Life with them outside the institution has always been much richer than life in front of the white board, life before the computer screen. Without these teachers—without Leverett and Judy, Tom and Susan; without Valerie; and without William and Andrew—no "education" would have been worthy of the name.

(Who's ready for a trip to the Pink Park?)

Table of Contents

CHAPTER ONE: INTRODUCTION

Problem One: Time

For English language arts teachers, the scarcity of *time* is among the most diabolical of rascals who plague us. And seemingly it has always been so. The earliest issue of the National Council of English Teacher's *English Journal*, for instance, features Hopkins' (1912) frustration with a recent effort to "apply the principle that pupils should learn to write by writing" (p. 2); the project in question had failed to account for the increased workloads of the study's teacher-participants, so that "without any material addition" to the number of teachers providing instruction and assessment, the study's only achievement was "merely a gratuitous increase in the labor of teachers who were already doing full duty" (p. 2). Hopkins' lament would serve equally well in today's deeply bifurcated teaching world. On one hand, strongly compelling voices continue arguing for the central place of writing across all academic curricula (e.g., National Commission on Writing, 2004, 2005; Applebee & Langer, 2006; National Writing Project & Nagin, 2006; Conley, 2007; Graham and Perin, 2007). On the other, Secretary of Education Arne Duncan has recently advised that the "new normal" of our current economic conditions may require us consider "smartly targeted increases in class size" (Klein, 2010; Sparks, 2010). Issues of personalized instruction and even mere classroom management aside—the additional time involved in educating a swollen roster of students is no small matter.

Part of the problem with ELA teachers' time is that so many competing demands attach themselves to it. This is not to say that the competing claimants are themselves inherently problematic. Some of these involve our legal and ethical commitments to the education of all students, regardless of whatever mitigating factors might require of them (and us) something other than an "ordinary" education (United States Department of Education, 2004). We all know

through experience the range of legitimate special expectations held by our students who have individualized needs because of giftedness, disability, belief, membership in a historically marginalized group, and so on. Addressing each expectation, however, requires time. Further claims on our time have arisen with the advent of the No Child Left Behind legislation (United States Department of Education, 2002), with its focus on certain core disciplinary knowledge and skills, and the arguably problematic testing regime it has spawned (Neill, 2003; Grobe & McCall, 2004; Cochran-Smith & Lytle, 2006; Darling-Hammond, 2006; Houston, 2007; McCarthey, 2008).

Still other time-takers are the result of our ever-increasing awareness that curriculum design and educational practices must be not only academically robust but also personally meaningful to our students. Thus, although the richly nuanced opportunities afford by a Deweyan approach to education (Dewey, 1910, 1915, 1938)—or even by its current reformulation in the work of methodologists like Wiggins & McTighe (2005)—can truly be said to be "efficient" from the learner's point of view (Dewey, 1915), orchestrating such opportunities requires an enormous investment of time from teachers, despite whatever comforting claims we might hear from those who would urge us never to work harder than our students (Jackson, 2009). Likewise, the hard-won enlightenment we are gaining against blindness to the historical embeddedness of our pedagogies (Counts, 1932; Freire, 1970/2000; Apple, 1979/2004) has brought with it the cost of the additional labor required of any educator who would teach "against" the text, against the *status quo*, against the grain of stultifying dogma.

Even when limiting our view strictly to commonly held ideal intended outcomes within the disciplinary framework of ELA, overt and covert demands on teacher time have increased rapidly over the past century. One way of understanding the increases can be seen in the

evolving characteristics of disciplinary conversations among educators in forums like *English Journal* – where an initial discourse heavily grounded in current-traditional approaches to the discipline (Connors, 1981a, 1981b, 1986) has slowly evolved into one more conversant with the process-based approaches of such sea-change thinkers as Emig (1971), Graves (1975), Murray (1982), and Atwell (1987/1998), and even of the socially and critically aware understandings of content-area specialists such as Nystrand (1997, 2006), Brandt (2001), and Purcell-Gates (2007). Simply put, as teachers have become aware that a heavily textbook-based approach cannot equal a more socially driven approach to learning, we have increasingly found ourselves abandoning memorization, drills, and workbooks—what Allington (2001) refers to as "stuff"—for more authentic, more engaging methods of instruction, practice, and assessment. But whenever we begin leveraging "knowledge" with authentic practices of reading and writing, with conversation in and out of the classroom, and with meaningful feedback to students about their progress, the ELA teacher's workload has a tendency to increase. Not sure? Pick a Saturday and drive by any high school that allows its teachers weekend access to the building. In the parking lot you're likely to find a disproportionate number of cars belonging to America's very best ELA practitioners, the ones whose robust lessons in reading, thinking, and writing are the most "efficient" in the Deweyan sense. For everything in life—as my AP United States History teacher, Mr. Washmon, used to say—there is a price to pay.

Perhaps the most taxing of our time-related operating costs involves ELA teachers' enormous investment of time in the practice of responding to student writing. "Grading" papers is, from one point of view, the bane of ELA teachers' existence. Sure, we endure under a regime of bells, and (many of us, gladly) serve the demands of supervisors who (rightly) demand of us that we use every moment of our fifty-minute hours to the educational advantage of our students.

Not only this, we also patiently bear up during our planning periods, which never get used for planning so much as for replying to emails and phone calls, entering grades, making photocopies, and meeting with students, parents, and problem-solving teams. But truly we *suffer* only when Friday rolls around, finding us in the process of stuffing our book bags for a weekend on the sofa, grading student essays. If you have ever stopped to think about it, you know already that the time commitment is staggering. But if you haven't, only a brief detour into basic math is necessary to understand how enormous the task can be.

Simply to read carefully (word-for-word, with slight circling back to clear up misreadings) a reasonably well crafted, two-page, MLA-formatted essay with its Times New Roman 12 font, one-inch margins, and double-spaced lines requires about three minutes of my attention. Perhaps I am a slower-than-average reader among my peer group, but that is the time it takes. I have approximately 140 student this year. If *each* of them composes a two-page essay for me to read and I read each essay carefully but without taking any time to make notations, I need 7 hours to complete my task. Of course, even a non-teacher realizes that my task involves much more than the act of reading itself. So suppose I allow myself approximately five additional minutes per paper—a number close to that suggested by the freshman English composition supervisor from whom I learned my first lessons in pedagogy—to provide feedback, assign a grade to each student essay, and record that grade in my online grade book.[1] Now my task has

[1] Although eight minutes per two-page paper sounds generous, it is nothing short of a mad dash. Moreover, it is a pace that I cannot sustain for more than about an hour at a time. In other words, the longer I grade, the slower I go . . . or the less feedback I give.

risen to one of 18.67 hours, over three seven-period days of class time.[2]

Even at this bare minimum of reading and evaluation time, each round of assigned essay evaluation requires of me and my colleagues somewhat more than three hours daily to avoid taking stacks of papers home each night and over the weekends. But given that teachers like me enjoy just under five hours of planning time in a contractual school week, and that much of this time is siphoned off by a host of other activities—planning lessons and assessments, reading course texts, making photocopies, responding to emails, reading texts, meeting with other faculty members, meeting with students or their parents, and so on—it is virtually impossible to grade even one set of papers weekly without having that work encroach deeply into what should be enjoyable time in the evenings and weekends spent with our families, our friends, and in the development of our own thought lives. Yet with that having been said, there is little hope of

---

[2] For comparison, see Sommers (1982). Writing in the context of college composition, Sommers comments, "More than any other enterprise in the teaching of writing, responding to and commenting on student writing consumes the largest proportion of our time. Most teachers estimate that it takes them at least 20 to 40 minutes to comment on an individual paper, and those 20 to 40 minutes times 20 students per class, times 8 papers, more or less, during the course of a semester add up to an enormous amount of time" (p. 148).

Of similar interest is Burkland and Grimm's opening passage in "Motivating through Responding," which captures well the existential angst of the teacher/evaluator:

> Faced with a stack of final drafts, many of us composition teachers prefer to clean the oven, pay the utility bills, or groom the collie. We play games to help ourselves through the task—"Five more tonight and I deserve a brandy before bed." Many of us find the hours spent writing response to final drafts to be the most time-consuming and most demanding mode of teaching. The fifteen to thirty minutes spent on one paper can mount to 23 to 45 hours for a teacher with a not unusual load of ninety students. These hours exhaust our heads and hearts . . . (p. 237)

helping our students improve their writing outcomes if we don't require them to write frequently or offer to them the opportunity of receiving rich feedback on much of what they compose.

Problem Two: Reliability

If time management weren't a complex enough issue in ELA teachers' response to writing, it is complicated by two other matters pertaining to the task itself. The first of these involves the institutional obligation we teachers have of making consistent judgments about the papers we receive, and of stratifying them according to their holistic successes against various sets of stipulated criteria. Peter Elbow has written about the challenge of reliable grading thus:

> For each essay in the stack, we have to decide between A, A-, B+, B, B-, C+, and so forth. If we use the full set of grades, we are using eleven levels (thirteen if we use A+ and D-). Even if we never use *any* grades below C-, we are still having to make fine evaluative discriminations among eight levels. . . . [Moreover,] we know that these decisions are not trustworthy, no matter how hard we agonize. Careful research has demonstrated over and over what common sense has told us—and what our students have learned through controlled experiments of submitting the same paper to different teachers: good teachers and evaluators routinely disagree about grades—and disagree widely. (p. 127)[3]

[3] See as well, for comparison, the following from Shaughnessy's *Errors & Expectations* (1977):

> Definitions of proficiency in writing vary widely from school to school and from teacher to teacher, with widest agreement at the lowest rung of the skills ladder, where correctness and basic readability are the concern, and the widest divergence at the upper rungs, where the stylistic preferences of teachers come into play. But even within the province of error, there are

Research into the causes and effects of these routine disagreements about grades, has a lengthy history in English education; unfortunately, it is a history more heavily concerned with guaranteeing the interrater reliability of large-scale measures such as the SAT, the AP Literature and English Exam, or perhaps even the Kansas Writing Assessment, than with the teacher-to-teacher comparisons about "fairness" that plague our students, their parents, and our administrators. Yet interestingly enough, the tradition of reliability studies actually began with a focus on the dysfunctional work of classroom teachers.

In the same year that found Hopkins complaining about researchers' failure to account for the amount of time involved in having students "learn to write by writing," Starch and Elliot published the brief paper "Reliability of the Grading of High-school Work in English" (1912). Prompted by the previous works of Dearborn (n.d.) and Jacoby (1910), Starch and Elliott had conducted a study in which two high school examination papers were distributed to two hundred high schools in the North Central Association "with the request that the principal teacher of first-year English grade these two papers according to the practices and standards of the school" (p. 449). The "startling" results demonstrated a "tremendously wide range of variation" (p. 454) far exceeding the ten-point range expected by the conventional wisdom of the day. In fact, the range produced by their study was "as large as 35 or 40 points" (p. 454). For obvious reasons, the authors were dismayed not only by the scores themselves and the variance they implied in real-world assessment, but also by the social consequences that would logically follow:

> For, after all, the marks or grades attached to a pupil's work are the tangible

---

> disagreements about the importance of different errors and about the number of errors an educated reader will tolerate without dismissing the writer as incompetent. (p. 276)

> measure of the result of his attainments, and constitute the chief basis for the determination of essential administrative problems of the school, such as transfer, promotion, retardation, elimination, and admission to higher institutions; to say nothing of the problem of the influence of these marks or grades upon the moral attitude of the pupil toward the school, education, and even life. (p. 442)

That the "promotion or retardation" of students was so greatly dependent upon "the subjective estimate of his teacher" was deplorable. Even worse was the realization that came in the next year, when Starch and Elliott published "Reliability of Grading Work in Mathematics" (1913), which demonstrated an "extremely wide variation of . . . grades even more forcibly than our study of English marks." So much for the charges of "subjectivity" in ELA grading, at least with respect to practices in the 1910s.

Among the studies that followed in the tracks of Starch and Elliot's early lead, many sought to illuminate reasons for the variation in teachers' scores. Marshall (1967), for example, prepared thirteen versions of a paper—one control, plus twelve variants demonstrating errors in spelling, grammar, punctuation, or a combination of the three—to show the degree to which readers devalued essays with formal errors even when directed to base their scores entirely on content. Distributing these instruments to 700 high school teachers of American history, Marshall found not only that scorers couldn't fully disentangle meritorious content from problematic form, but also that errors in spelling and grammar accounted for lower grades than did errors in punctuation. Further, the study demonstrated that the combined-error papers with the greatest number of errors were scored *less harshly* than those with only a moderate number of spelling- or grammar-only errors. Marshall supposed that spelling and simple grammar errors were easy to spot and that they provoked unconscious conclusions about the student's overall

ability, despite the fact that these sorts of errors were the least likely to create reader confusion. Whatever the causes, Marshall's test implied that even when given a rubric and a model paper, real-world evaluators were unable to provide reliable scorings of written work.

Soon thereafter, the effect of handwriting on essay scores was also demonstrated. Chase (1968) found that readers scored poorly scripted versions of essays lower than well scripted ones, particularly whenever a "negative halo" pertaining to decoding a first essay asserted itself over a second one. In his study, readers evaluated two essay samples. When faced with poorly scripted samples, they initially ignored the difficulties associated with the script itself, yet by the second sample their scores dropped considerably. Marshall and Powers (1969) conducted a similar study to illuminate possible interactions between handwriting neatness and compositional errors. While they found no such interactions, two surprising results did emerge. First, "neat, easy-to-read, handwritten" essays outperformed content-equivalent typed versions. Second, the highest-to-lowest ordering of mean scores for each written form was as follows: "neat" (5.66, S.D. 1.62), "poor" (5.25, S.D. 1.63), "typed" (5.15, S.D. 1.71), and "fair" (5.02, S.D. 1.57). The authors were at a loss to determine whether the study had turned up "an artifact of the somewhat unusual grading situation, or whether it was a reflection of the actual effects of writing neatness on essay grades" (100).

Diederich's *Measuring Growth in English* (1974) provided further evidence of evaluator variance in a discussion of work he, John French, and Sydell Carlton completed for ETS in 1961.[4] For their study, the researchers obtained 300 essays written by students of three colleges.

[4] A more colorful expression of Diederich's take on the low reliability of essay ratings appears in the proceedings of NCTE's 1963 national conference: "I honestly believe that almost all experiments concerning

Then, to determine "what qualities in student writing intelligent, educated people notice and emphasize when they are free to grade as they like," they distributed identical copies of all 300 papers to each of 60 readers from across six professional fields—college instructors of English, social sciences, and natural sciences; writers and editors; lawyers; and business executives. None of the graders communicated with each other, nor were they given any rubric other than the instructions to "sort the papers into nine piles in order of general merit, using their own idea of what constituted general merit" (p. 5). The graders were obligated to use all nine piles, with no fewer than 12 papers per pile. They were also asked to make brief comments about strengths and weaknesses "on as many papers as possible" (p. 5). 53 of the 60 readers completed their tasks. The results of this study were astounding:

> The reliability of grading that was shown in this study should not be taken to represent the reliability usually attained in grading essays for the College Board, when we adopt strict rules and enforce them by close supervision. But it is probably typical of the amount of disagreement one would find in any large group of readers without such training and discipline that, out of the 300 essays graded, 101 received every grade from 1 to 9; 94 percent received either seven, eight, or nine different grades; and no essay received less than five different grades from these fifty-three readers. (p. 6)

Diederich, whose experience as an Educational Testing Service researcher had led him "to accept a reliability of .80 in the measure . . . of an important objective as adequate for practical

---

English composition that rely on essay grades have been conducted with tape measures printed in elastic" (Diederich, 1964, p. 60).

decisions in the ordinary course of schoolwork" (p. 2),[5] found in this study that the median individual-to-group grader correlation was a mere .31.

Following Diederich, studies in interrater reliability continued looking for reasons to explain the variance that researchers had uncovered. Freedman (1979), for example, completed a study to determine the various effects four domains of success—content, organization, sentence structure, and mechanics. To do so, she began with a set of moderately well-composed college essays written on eight topics, which she then rewrote to strengthen or weaken their outcomes in the four domains. She then distributed these variants to twelve readers who had been recommended as experts by their colleagues at Stanford University. Prior to their scoring task, these readers received training on a 4-point holistic rubric by means of a set of practice essays. When grading, they were asked to supplement their scores with a detailed commentary regarding content, organization, sentence structures, and mechanics. Reliability among the scores was high, between .86 and .96, but Freedman conceded that the extreme differences in the essays themselves may have accounted for such agreement. An ANOVA of the results determined that differences in content provided the largest main effect (a 1.06-point difference on a 4-point scale between strong- and weak-content papers), followed by differences in organization (nearly a 1-point difference) and mechanics (½-point difference). Freedman interpreted these results to signify that her raters were not as attuned to sentence structure and mechanics as to content and organization.

Freedman closed her article with a series of useful critiques directed toward the

---

[5] Hillocks (1986, p. 101) also implies .8 as an acceptable level of reliability in the scoring of written samples.

profession. First, if—as her study implied—society holds content and organization as more important than sentence structure and mechanics, teachers "should aim first to help students develop their ideas logically" and then to "focus on teaching students to organize [these] developed ideas" (p. 336). Indeed, considerations of organization should be taught "before or at least alongside those of mechanics and sentence structure" (p. 336). Second, even if they ignore this advice, teachers should avoid making claims of valuing content and organization while providing comments focused more heavily on mechanics. Finally, the profession as a whole might improve writing instruction by "understanding how evaluators evaluate as they do" (p. 337).

By the 1980s, researchers like Chase (1983) were noting common denominators among the elements that interfered with reliable scoring. Chase writes, "They all involved variables that complicate the processes of reading the essay. To the extent that the reader must concentrate more the decoding of the writing, he or she may attend less to the content of what has been written or may transfer frustration in decoding to lower marks for the paper" (p. 293) Considering this observation, Chase hypothesized that "any condition that complicates readability should reduce scores on essays" (p. 293). And so he set out to prove this theory by having readers evaluate alternate versions of a content-identical essay—one at an "easy" reading level, the other more challenging. The study's readers were master's and doctoral students in educational measurement classes who had recently studied the topic of construct validity—the topic of the essay itself. Each reader received one essay and the instructions to grade it solely on the correctness and development of its information, without consideration of any other factor. Not surprisingly, the more difficult-to-read essay received lower scores. Chase inferred from his results two possible rationales for the lower grades. Perhaps the challenges of decoding a text

were distracting readers from grasping its content. Alternatively, readers may simply have a threshold of "reading difficulty they will readily accommodate" (p. 296) before taking out their frustration upon an essay's score. In a subsequent study, Chase (1986) demonstrated not only that matters of readability affect scores, but also a host of interacting variables, including "the reader's achievement expectations for the student, the sex and race of the student, and the quality of penmanship all have an effect on the score given a child's essay test" (p. 40).Yet despite these and other observed challenges to the reliable scoring of essays, the obligation of grading consistently is still (and presumably always will be) one of ELA teachers' central job targets.

Moreover, despite a concurrent set of concerns about the validity of large-scale writing assessments (White, 1995; Huot, 1996), the perceived need for consistently scored standardized written examinations has generated an entire industry devoted to producing several ubiquitous series of high-stakes tests whose very existence depends upon their rather high degrees of reliability. Two of these, the College Board's Advanced Placement English Language and Composition and its Advanced Placement English Literature and Composition exams, boast reliability scores of .758 and .805, respectively, in recent studies (College Board, 2007). Likewise, College Board's SAT essay, while only achieving correlations in the mid 50s (e.g., Pearson = .56, Coefficient Alpha = .55), nevertheless can make the claim that "for the average student who scored in the 6 to 7 range, well over half can expect to score within one point of their initial score, about one-fourth can expect an increase of 2 to 3 points, and about one-eighth can expect a decrease of 2 to 3 points" (Breland et al, 2004). Presumably, because the essay portion of the SAT is combined with a multiple-choice component for a final score, students' total verbal scores have an even higher reliability.

High reliability is not just an ideal for large-scale, high-stakes testing. It is, in fact, one

with which classroom teachers should struggle to improve. Wiggins and McTighe (2005) note two fundamental means by which teachers may accomplish such improvement. First, teachers should build into their assessment procedures a series of "multiple tasks for the same outcome," as "better reliability is obtained when the student has many tasks, not just one" (p. 348). Second, teachers would benefit from remembering that "scoring reliability is greatly improved when evaluation is performed by well-trained and supervised judges, working from clear rubrics and specific anchor papers or performances" (p. 348), procedures not unlike those used by large-scale testing organizations.

In addition to multiple measures, rubrics, and anchor pieces, teachers might consider professional development strategies as another hedge against inconsistent scoring. One promising approach was attempted and discussed in the early 1900s by the freshman composition group at University of Illinois. Tieje, Sutcliffe, Hillebrand, and Buchen (1915) report their department's response to the problem of fairness for the 1450 students of a program taught by 25 different instructors.[6] As reported in the authors' discussion, the staff of Rhetoric I

[6] Although the current applicability of an assessment procedure used by a major university a century one century ago could be initially perceived as of limited value, Tieje, et al., captured my attention because of their group size. Similarly to the program described in this essay, my high school's English language arts department serves approximately 1600 students with a staff of 16 full- and part-time teachers. Thus the factors contributing to concern over reliability at the University of Illinois are of nearly the same magnitude as I find in my own department. Moreover, the Professional Learning Communities model (DuFour & Eaker, 1998; DuFour, DuFour, Eaker, & Karhanek, 2004), which has over the last decade put rather deep roots into my school's culture—operates by principles not unlike those described by Tieje, et al. In short, not only do I find this piece intriguing as a historical

committed themselves to a standard schema for assessing student work, bearing the following qualities. First, the schema was built upon the aim that "the first semester's work in composition . . . must be to remove such traces of illiteracy as still remain, and at the same time to give some advanced instruction in the principles of composition which shall enable the student write unified and coherent, if not emphatic, exposition" (p. 590). In part this aim would be accomplished by the writing and assessment of essays, but also in part through the assignment and collection of exercises in a composition handbook. Second, instructors were to grade essays by a fixed, rather than sliding, standard. Although a sliding standard would allow for "the development of the student and for the acquisition of new facts of rhetoric in the course of instruction" (p. 588), it would present too great a challenge for uniform implementation by so many different instructors. Third—and perhaps most relevant to the ongoing question of consistent grading—the instructors were to devote time in their weekly meetings to the grading of a model essay and to discuss the reasons for the grades given, "with the hope of obtaining uniformity" (p. 587) in their assessed (i.e., numerical) values.

The matter of consistent scoring within and across classrooms is an important one. Students deserve to experience enough consistency in grading from each assessment to the next—and from each teacher to the next—that they may adequately understand where they stand in relation to their school systems' expectations. But that having been said, strong consistency in grading among classroom teachers will almost certainly prove to be an elusive target. Much as noted in the Diederich's (1974) study, ELA teachers are unlikely to find themselves being

---

account of educational problem-solving, it strikes me as an interesting direction for new research in embedded, ongoing professional development models directed at increasing interrater reliability.

required to "adopt strict rules [about within-classroom scoring] and enforce them by close supervision" (p. 6). Moreover, they will continue to face multiple threats to uniformity in response, among which are variance in the evaluators themselves, their abilities, backgrounds, training, and general dispositions; variance in their perceptions of the task and its proper judgment; variance in their idiosyncratic interactions with the various components of students' individual performances, as when some but not all teachers become hypercritical about sentences that begin with coordinating conjunctions or end with prepositions, or when they respond differently to matters of spelling and vocabulary selection, or when their readings tend to focus too exclusively in the direction of content, or organization, or conventions, or any other distinguishable feature; and even variance pertaining to the contexts in which teachers provide their assessment, as when they are laboring late at night against deadlines, struggling in February with the depth of their work loads, or returning to grading after a summer of dormancy. As my district's coordinating teacher for ELA has reminded me, teachers grading in May is substantially different from their grading in August.

And all of these sources of variance—what my ANOVA professor refers to as "noise"—have the very real potential of overpowering our ability to pick up on the signal of true variance. So it is of no wonder that Elbow or others might find the level of agreement in teacher grading to be so low as to be "not trustworthy." Nevertheless, if instead of giving up on the problem as hopeless, we could find our way—as the University of Illinois Rhetoric I group seemed to have done in the early twentieth century—to any increase in interrater reliability among a contextually similar teaching faculty, we would undoubtedly be doing our students a good service. For grades—no matter what we may have shown or believe we know about their benefits or harms (e.g., Harter, 1978; Butler & Nisan, 1986; Grolnick & Ryan, 1987; Kohn, 1994; Pedersen &

Williams, 2004; Kitchen, et al, 2006; O'Connor, 2007)—are among the institutional ways that we communicate with students about their relative levels of academic success. If we are to offer grades at all—which, currently, we must—they should be as reliable as possible.[7]

Problem Three: Relevance

Although the pedagogical value of reliable scoring—the instructional relevance of giving consistent grades—is a matter of ongoing debate, its central place in traditional educational systems is at once both a rationale for our continued attention to more reliable grading practices and also a potential threat to the sort of feedback that is likely help our students make their best progress as thinkers and writers. For if we take the grading (i.e., the sorting) part of our work seriously, we must call upon ourselves not only to be accurate judges of what separates one work from another but also to communicate what constitutes the bases of our judgments—to offer what I will refer to as *sorting-oriented* feedback.

Sorting-oriented feedback, as I will use the term, involves any commentary whose central purpose lies in justifying the grades given to a text (Dohrer, 1991; Elbow 1997) rather than in

---

[7] As a side note, I see much wisdom in the philosophical and research positions of those who believe that students' focus on achievement for the sake of grades can produce the tendency for revisions to aim at improved scores rather than improved writing (see Kohn, 1994 for a brief synopsis). I am also aware of studies showing that grades themselves have a negative effect on students' outcomes over time (e.g., Butler & Nisan, 1986). But I wonder if there aren't some missing pieces in this line of research. I wonder, in fact, if subsequent research might not show that reliable grades *in combination with* near-optimal feedback and an open-ended policy for revisions might produce better results that comments-only feedback on formative drafts followed by a final "graded" draft for which there is no redress. The work of Kitchen, et al (2006) would seem to indicate *no*, but I am intrigued enough that I might follow this line a little further in future work.

provoking thoughtful reflection from students before they return to improving their work. Perhaps in this it represents a tangible "way for teachers to satisfy themselves that they have done their jobs" (Sommers, 1982, p. 155), perhaps too often in a manner that focuses more on a text's deficiencies, its "formal and technical flaws," than on its "intended meanings," thereby diminishing students' "incentive to write" and their "motivation to improve skills" (Brannon & Knoblauch, 1982, p. 165). Sorting-oriented feedback may thus be something of a symbolic hand-washing ritual, offered by teachers in lieu of further engagement in richly meaningful dialogue about improving texts. At its worst extreme, sorting-oriented feedback can be not only dismissive but actually go so far as to be interpretable as "manifesting scorn, hostility, condescension, flippancy, superficiality, or boredom" (Horvath, 1984) to our student writers.

Unfortunately, in a grades-based world, sorting-oriented feedback is something of a necessary evil—an activity that is to educators what defensive medicine is to physicians (American College of Emergency Physicians, 2011; Gore & Lloyd, 2011), a cover-your-ass move often meant to steal the thunder from students (and their parents) who aren't happy with what a particular grade might be doing to their course average, to their GPA, to their opportunities for placement into the right college, or to their likelihood of receiving a merit-based scholarship. The need for sorting-oriented feedback hovers in the back of any teacher's mind who has received emails like one sent to me this spring from a disgruntled mom: "If this [grade] in any way reflects on [my daughter's] records for college I will be speaking to the administrators of the school."

Sorting-oriented feedback may be an institutional necessity; it is undoubtedly an energy-draining chore. Providing reliable sorting-oriented feedback demands that we repeatedly exercise judgments about the same construct of interest, over and over again. When wearing our

"reliability hats," it seems that we reduce the field of what we can observe and report to the qualities that can easily be stratified or categorized: *Did you provide a correct MLA heading or not? How precise are your margins. How many errors can I find in your Works Cited list? How many comma splices have you failed to correct?*[8] Ask any teacher of writing; this is not happy work. Moreover, it operates from an inherently antagonistic stance. It begins with the implicit (or explicit, if codified into a rubric) promise that *I will be taking away points whenever I encounter [X]*—a promise that rings true even for ostensibly "rewards-driven" rubrics like that used by Advanced Placement, with its statement that students are to be "rewarded for what they do well" (College Board, 2010). Despite such happy language, the AP rubric like any other is about finding reasons to sort students into various categories of achievement.

Yet even sorting-oriented feedback does have a benefit in that it helps students understand clearly and specifically where their trouble spots lie (Dohrer, 1991; Lynch & Klemans, 1978, Sommers, 1982; Land & Evans, 1987; Straub, 1997), or what sorts of solutions might be in order (Straub, 1997). Without it as a bare minimum of commentary, by contrast,

---

[8] In this and the surrounding points, the framework for sorting-oriented feedback has much in common with the sorting orientation of standardized writing assessments in general, as expressed by Condon (2009):

> [A]s writing assessment is practiced more often than not, it is an essentially reductive enterprise. Because the goal is to reach a score or a ranking that will assist in making an placement, and because those placements are sufficiently high-stakes to necessitate close attention to validity (in all its manifestations) and reliability, we reduce the construct *writing* to only those parts of writing that are obviously measurable, we carefully train raters to attend to only those factors, and we pretend that the varied set of competencies that combine to produce "good writing" can be expressed in a single number. (p. 141).

grades would undoubtedly seem to students a wholly alchemical, wildly subjective sort of feedback. Indeed, because of society's perceived value in grading, the sorting-oriented feedback attached to it neither may be avoided by classroom practitioners nor ought it be suppressed by well-meaning researchers whose concerns for social justice rightly challenge the culturally stratifying (i.e., class reproducing) effects of grades. What we need instead are ways to supplement our sorting-oriented feedback with generous amounts of what I will refer to as *learning-oriented* feedback.

Where sorting-oriented feedback runs the risk of being the door-closing defense of a grade, the demonstration of deficiencies or flaws, or the signal of a stopping point beyond which further instruction and revision will no longer take place with a current work, learning-oriented feedback by contrast operates as a clear, open invitation to further learning. It is a best-practices sort of pedagogical communication rising to a challenge well expressed by Sommers (1982):

> The challenge we face as teachers is to develop comments which will provide an inherent reason for students to revise; it is a sense of revision as discovery, as a repeated process of beginning again, as starting out new, that our students have not learned. We need to show our students how to seek, in the possibility of revision, the dissonances of discovery—to show them through our comments why new choices would positively change their texts, and thus to show them the potential for development implicit in their own writing. (p. 156)

In helping students arrive at moments in which the "dissonances of discovery" can occur, learning-oriented feedback presses beyond the mere illumination of error. Again Sommers (1982) arrives at the heart of what is accomplished when teachers offer richly meaningful, learning-oriented feedback:

> Instead of finding errors or showing students how to patch up parts of their texts, we need to sabotage our students' conviction that the drafts they have written are complete and coherent. Our comments need to offer students revision tasks of a different order of complexity and sophistication from the ones they themselves identify, by forcing students back into the chaos, back to the point where they are shaping and restructuring their meaning. (p. 154)

As will be discussed more fully in chapter two, optimal feedback is among the most powerful drivers of growth in all of teaching and learning, with an effect size of 0.79 or nearly twice that of school in general (Hattie & Timperley, 2007). And as will be discussed somewhat more fully in chapter two, optimal feedback possesses the following content characteristics (Brookhart, 2008):

- It maintains a *focus* is on the student's work itself, on the processes used by the student to complete this work, or on the student's self-regulatory processes; it avoids a focus on the student individually as a person.
- It makes *comparisons* either to the criteria for "success" or to the student's prior performances.
- It adopts a *function* of describing rather than judging the student's work, processes, or self-regulation.
- It maintains a positive *valence*, either by drawing attention to what has been done well or—when needing to point out difficulties—by noting not only the "errors" but also suggested avenues for improvement.
- It achieves *clarity* in communication by the use of developmentally appropriate vocabulary and concepts, and by offering an amount of comments that is useful

but not overwhelming.

- It attains *specificity* in the comments so that students can envision precisely what their next steps might be.
- It develops a respectful *tone* that positions the student as the task's agent and provokes in the student a desire to continue thinking about the task as one still-in-process.

Achieving such traits while carrying a full course load of 100 or more students may be something of a career-long challenge. Nevertheless, to the degree that we aspire to these, our feedback has the potential to nurture a predictable constellation of beneficent outcomes, among which are:

- Reinforcing classroom instruction (Sommers, 1982).
- Inducing richer understandings about what good writing looks like (Sommers, 1982).
- Increasing the likelihood of producing risk-taking as opposed to mere error-avoidance (Horvath, 1984).
- Strengthening our students' sense of their own control over their writing (Brannon & Knoblauch, 1982), knowing that when we do so, we "create a rich ground for nurturing skills because the writer's motive for developing them lies in the realization that an intended reader is willing to take the writer's meaning seriously," that because "the writer is allowed to have something to say . . . the saying of it is more likely to matter" (p. 165).
- Building better working relationships with our students (Straub, 1996).

In short, learning-oriented feedback is a fundamental part of the ongoing dialogue between

teacher and student, a way of communicating that is likely to deepen our students' engagement with their selected topics or with their "purposes and goals in writing a specific text" (Sommers, 1982, p. 154; see, also, Freedman, 1987).

If—with such good opportunities for teaching and learning hanging in the balance—we are to transcend the transactional utterances of sorting-oriented feedback, we must be careful as responders to develop such characteristics in our comments. This is to say that we must be guided by important self-reflective questions about the comments we put to page. *To whom will our notations be relevant, and for what purposes? To what degree can we foster situations in which a larger proportion of our time is wrapped up in an approach to the sort of commentary that drives learning?* Returning once again to Sommers (1982), we will do well to remember the end goals of our practices:

> We comment on student writing because we believe that it is necessary for us to offer assistance to student writers when they are in the process of composing a text, rather than after the text has been completed. Comments create the motive for doing something different in the next draft; thoughtful comments create the motive for revising. Without comments from their teachers or from their peers, student writers will revise in a consistently narrow and predictable way. Without comments from readers, students assume that their writing has communicated their meaning and perceive no need for revising the substance of their text." (p. 149)

To Sommers, I would only add one further thought. *To the degree that we become writers of commentary that is relevant for our students as learners, we will also become responders who find in our work of essay evaluations one of the least burdensome, least exhausting parts of our*

teaching routines. The ELA teacher's enormous workload may be a constant, but there is no need for its most significant component also to be inherently wearisome or frustrating. In fact, it should be a rather joyful task.

And thus the problems in which I'm currently most interested have come to focus: While grades must be given to satisfy teachers' institutional demands for ranking and sorting students, and the justification of grades is a necessary evil in our work of assessment, teachers need a better apparatus to shrink the amount of time needed in the suggestion and defense of grades, *per se,* so that we may better use our limited time in offering to students more of what actually drives their learning: *meaningful feedback.* And if we can do so in a way that neither succumbs to grade inflation or lowered interrater reliability, all the better.

Purposes of the Study

Even when we aren't particularly careful about it, feedback has a tendency toward positively impactful results (Kluger & DeNisi, 1996). When we adhere to "best practices" in feedback, however, our efforts are twice as impactful as school in general (Hattie & Timperley, 2007). Given what we know about the characteristics of particularly beneficial feedback, this study has proposed to test the outcomes of a simulated in-class activity to discover on one hand whether it has the tendency to increase teachers' rates of feedback that pushes beyond attention to superficial matters of "correctness" toward a more conversational realm involving not only the paper itself, but also the processes the student has used to complete the work, the student's sense of self-regulation, and the student's observance of the possibilities of change from one draft to the next; and on the other hand whether the feedback takes on a richer, more positively helpful tone in its advice.

The in-class activity in question comprises the following elements:

- Students arrive to class with complete, ready-for-submission drafts, drafts traditionally considered "final."
- The teacher provides students with copies of the rubric by which their work will be judged, supplementing this rubric with illuminating examples on an overhead or digital projector. The examples may demonstrate "correct" solutions regarding the skills in question, as in the case of MLA formatting; they may also demonstrate common problems as well as reasonable "fixes" for errors such as those pertaining to punctuation, passive voice, and the like.
- Students receive instructions to examine the rubrics, the examples, and their own papers, looking for places at which they have achieved or failed to achieve the assignment's particular learning targets. Where they find problems in their texts, they should *correct these by hand*.
- Students then use their rubrics and experience in the class to predict the scores their essays should receive.
- Finally, students write as many as two specific questions about which they are most interested in receiving targeted feedback from the teacher.
- In exchange for the students' careful attention to these matters, the teacher will count any last-minute corrections as though they were already part of the ready-for-submission drafts that students brought to class. Further, the teacher will begin the feedback-providing task by giving careful attention to the questions specifically posed by the students themselves.

Because part of the student activity involves a predicted score and because teachers' knowledge of this predicted score—as well as their "distraction" stemming from the student

comments—might have the effect of altering their grade-wise perception of the student essays, the study has also sought to investigate any possible changes the in-class activity might provoke with respect to interrater reliability (as measured by intraclass correlation) and to the averages and distributions of grades assigned to the essays being evaluated by the study's feedback-offering teachers.

Importance of the Study

Providing rich feedback to students' written work is a massively time-consuming, core component of the ELA teacher's professional obligations. Yet the amount of instruction we receive in composing such feedback is minimal to nonexistent in our teacher preparation programs. As a new teacher in the early 2000s, I arrived on campus with a comparatively enormous background in assessment and feedback, having worked for two years as a university-paid writing tutor and one year as the teacher-of-record for four sections of freshman composition. But even with this, I had received perhaps much less than five hours of total instruction in offering feedback—all of it within the context of my work as a university classroom instructor, none of it in the state-accredited teacher certification program through which I earned my credentials as a "highly-qualified" ELA teacher. My experience is far from unique. As an adjunct general methods instructor at a greater Kansas City-area university during 2010-2011, I have witnessed a similar paucity in emphasis that assessment and feedback have received in my own students' courses of study. And in conversations around the lunch-table with my high school colleagues, I've learned that their own preparatory experiences are much like my

own. Poor.[9]

Yet despite such backgrounds in teacher training, my belief is not so much that that experienced teachers are intellectually (or at least intuitively) unaware of the benefits to be obtained by communicating richly to students our beliefs about their written performances, nor that teachers are uninterested in providing such feedback. Rather, I would submit that many teachers feel free to offer learning-oriented feedback only after having satisfied the perceived sorting-oriented obligation laid upon them as institutional gate-keepers. In other words—and in keeping with the current culture's perceived attacks on the profession—we teachers spend too

---

[9] This lack of background should not be surprising, given statements such as the following by B. Huot about the lack of interest in assessment theory and practice for higher-education composition programs:

> I think of composition as sort of the Rodney Dangerfield of the academy, and then I see assessment as the Rodney Dangerfield of composition. . . . It really is. It's something that nobody likes. We don't want to talk about it. If someone works in that area, they're automatically suspect. That may be changing—I mean, I'm not on the market but I have students who are, so I've seen some of the positions being advertised and it seems to me that there are quite a few jobs these days that are looking for people to have expertise in assessment. In fact, somebody called me recently wanting a more senior person, but there are only a handful of senior people that I can even think of who even work in assessment. So I think that we need to try to rehabilitate assessment if we possibly can because I think it's really important. (Bowman, Mahon, & Pogell, 2004)

It is worth noting, moreover, that methods courses in English education programs frequently "appear to present only the most general knowledge about writing, focusing instead on literature," and that even in this, "courses devoted to writing tended to be workshops for students to work on their own writing rather than courses in the teaching of writing" (Smagorinsky and Whiting, 1995, p. 74; see, also, Kennedy, 1998). In such a context, it completely understandable that the *assessment* of writing receives very little attention.

much of our time providing defensive feedback rather than engaging students more through more meaningful types of commentary because we feel we must. And no matter how much we might wish to do so, relaxing from our a defensive postures and embrace a wholly learning-oriented approach isn't exactly possible, either. Sorting-oriented feedback, just like the grading practices it supports, is part of the river in which we swim. And in practice it matters very little whether we chafe against the hegemonic role that grading plays in our educational systems, for even in the best imaginable of post-NCLB eras grading is not likely to simply go away.

The current study may prove important, then, in multiple ways. First, it has been conducted within the framework of offering a professional development session to middle- and high school teachers, the majority of whom most likely possess assessment training backgrounds not much more rich in the area of providing feedback than my own. For them, mere participation in the study may have served as a reminder of (or alert to) the rich possibilities in feedback that may not be part of their present practices. Second, the study may have proven important in that it has suggested to its participating teachers a procedure that *by design* is intended to facilitate a more comprehensive feedback approach because it delegates to students at least some of the sorting-oriented responsibilities to which feedback usually responds. With such suggestions having been made, it would not be surprising if a few were to awaken to the same realization expressed in Fuller's "Teacher Commentary That Communicates: Practicing What We Preach in Writing Class" (1987): "What my commentary did was not communicate to a person but make marks on a text, marks that were random and disparate criticisms of the formal properties of a text; in effect, notes to a paper, not response to a writer" (p. 308). To the degree that this study may have helped its participants understand that they may partially delegate their sorting-oriented obligations to students themselves and redirect their own attention to the *writers behind*

texts, it will have succeeded at cultivating responses to *people* rather than merely to *papers*, no matter the hypotheses' outcomes.

Finally, the study may have proven important to the degree it has supported the truth of its hypotheses. In short—given how much time I and my departmental colleagues invest in providing written feedback to our students—nothing would be much more remarkable in the day-to-day world of our work as high school ELA teachers than that we had found a dependable method of communicating with our students more richly about their growth as writers while both diminishing the amount of time devoted to sorting-oriented feedback and also holding constant the reliability of the grades we are obligated to affix to their final submissions. It is thus my hope that the study has gone so far as to demonstrate a means by which ELA teachers can reallocate the time we devote to providing comments about students' work—sacrificing less energy toward the rather adversarial task of sorting papers by their outcomes so that we might re-invest it into the more pedagogically meaningful task of learning-oriented feedback. At least two reasons suggest this as a meaningful course of action. On one hand, well considered feedback has repeatedly been to be a major engine for students' academic growth. On the other hand, the professional and personal satisfaction we teachers might gain from interacting with students about their ideas and outcomes would far outweigh the rewards of sorting papers into piles of good, bad, and indifferent success. An alteration in our approach to feedback might thus come to be seen as beneficial both to our students and to ourselves as well.

Theoretical Framework

Hattie and Timperley's metastudy "The Power of Feedback" (2007) outlines four levels of focus to which feedback can attend: *focus on the task (FT)* or product submitted by the student, its correctness, formal features, and the like; *focus on the processing used to complete*

the task (FP), such as those the student would need to understand or accomplish in order to better achieve desired outcomes for the task; *focus on the student's self-regulation (FR)*, perhaps regarding the student's ability to self-evaluate the need to understand or execute better, the student's self-efficacy or ability to self-regulate; and *focus on the student personally, the student's "self" (FS)*, apart from specifically identifiable interactions between the self and task, processing, or self-regulation. FT, FP, and FR have been shown repeatedly to be of benefit to academic learners. The effects of FS, however, "are too diluted, too often uninformative about performing the task, and too influenced by students' self-concept to be effect. The information has too little value to result in learning gains" (p. 96).

Brookhart (2008) has adopted these four levels of focus, drawing as well from the research of Bangert-Drowns, Kulik, Kulik, & Morgan (1991); Butler & Winne (1995), Kluger & DeNisi (1996) and others to formulate for classroom teachers a multidimensional rubric of characteristics for *feedback content*, as reproduced in Figure 1. Brookhart proposes, based on the research she has reviewed, that teachers should engage in FT, FP, and FR extensively—FS rarely if at all; that they should offer *criterion*- and *self-referenced* comparisons to student outcomes, but generally avoid *norm-referenced* comparisons to other students' works, in that such a comparative mode "creates winners and losers and plays into that fatalistic mind-set that says student ability, not strategic work, is what's important" (p. 23); that they should offer the majority of their comments as *descriptions* rather than *judgments* (including the judgments implied by grades themselves); that they should be *positive* in describing the student's achievement of criteria or at least by offering "things the student could do about it" (p. 26) where criteria have been missed; "[j]ust noticing what is wrong without offering suggestions to make it right," says Brookhart, "is not helpful" (p. 26); and that their comments be clearly

understandable, given with reference to specific locations within the text, and offered in an unequivocally helpful tone that respects the student as a self-efficacious agent of her own education.

Figure 1 Feedback Content (Brookhart, 2008. pp. 6-7)

Feedback Content Can Vary In . . .	In These Ways . . .	Recommendations for Good Feedback
Focus	• On the work itself • On the process the student used to do the work • On the student's self-regulation • On the student personally	• When possible, describe both the work and the process—and their relationship. • Comment on the student's self-regulation if the comment will foster self-efficacy. • Avoid personal comments.
Comparison	• To criteria for good work (criterion-referenced) • To other students (norm-referenced) • To student's own past performance (self-referenced)	• Use criterion-referenced feedback for giving information about the work itself. • Use norm-referenced feedback for giving information about student processes or effort. • Use self-referenced feedback for unsuccessful learners who need to see the progress they are making, not how far they are from the goal.
Function	• Description • Evaluation/judgment	• Describe. • Don't judge.
Valence	• Positive • Negative	• Use positive comments that describe *what* is well done. • Accompany negative descriptions of the work with positive suggestions for its improvement.
Clarity	• Clear to the student • Unclear	• Use vocabulary and concepts the student will understand. • Tailor the amount and content of feedback to the student's developmental level.
Specificity	• Nitpicky • Just right • Overly general	• Tailor the degree of specificity to the student and the task. • Make feedback specific enough so that students will know what to do but not so specific that it's done for them. • Identify errors or types of errors, but avoid correcting every one (e.g., copyediting or supplying right answers), which doesn't leave students anything to do.

Tone	• Implications • What the student will "hear"	• Choose words that communicate respect for the student and the work. • Choose words that position the student as the agent. • Choose words that cause students to think or wonder.

Research Questions

The current study has proposed to leverage previous research in effective feedback practices in order to demonstrate the worth of a self-feedback routine administered to students prior to their submission of final drafts for teacher evaluation. It is believed that that the self-feedback routine will precipitate three sets of related consequences with respect to teacher's evaluative practices: (H1) a mild inflation of the general-impression grades assigned to student work, perhaps best explained as an inflation resulting from some teachers' attention being diverted away from clusters of easily spotted and disproportionally penalized mechanical and conventional writing errors; (H2) a marked improvement in the agreement of these general-impression grades—as measured by intraclass correlation—as the teachers who would normally penalize basic errors close ranks with those who observe such errors through kinder evaluative lenses, doing so because they have been influenced by the students' corrections of routine mistakes as well, perhaps, by the students' own assessments of outcomes; and (H3) a dramatic increase in the overall richness of their feedback against the criteria stipulated by Brookhart (2008). With respect to this third hypothesis, several subsets are to be observed:

- $H3_A$: Although FT and FS will remain proportionally constant across experimental conditions and paper strength, the proportion of comments focused on the student's composing process (FP) and self-regulation (FR) will increase under the experimental condition—and more notably so on the weaker papers—provoked by the student's own handwritten self-evaluative comments having been

added to the word-processed essays.

- $H3_B$: Comparisons to the criteria for "good writing" will remain proportionally constant across experimental conditions and relative paper strengths, but comparisons to imagined previous and/or successive drafts will increase under the experimental condition—and more notably so for weaker than stronger papers. Comparisons to the norm of other students' work will be minimal and constant across both groups, as teachers will not have access to enough representative texts to form concrete notions about group norms.
- $H3_C$: The proportional amounts of descriptive and evaluative comments will remain constant across experimental conditions and relative paper strengths, as teachers' responses are likely to be similarly descriptive or evaluative regardless of whether they are responding to the student's text per se or to the student's comments about that text.
- $H3_D$: A higher proportion of comments will possess positive valence in the experimental condition and with higher-quality papers, as teachers in both situations will adopt a model of communication best described as evaluator-to-person rather than evaluator-to-text. This is to say that as teachers respond to better papers and to papers supplemented with student-provided commentaries under the experimental condition, they will more frequently rise above mere valence-neutral language of editorial symbols and simple edits, and into domains of communication that involve a more interpersonally "positive" and engaging manner of describing the text's strengths and weakness.
- $H3_E$: The proportions of comments judged to be "specific" or "unspecific" will

remain constant across conditions and degrees of paper strength.

- $H3_F$: The proportions of comments judged to be "specific" will remain constant across conditions.
- $H3_G$: As with valence—a measure of "positive" communication, even when communicating the necessary improvements to a text—the proportion of comments judged to be helpful in tone (respectful, positioning the student as agent) will be greater in the experimental condition and with stronger papers.

Apart from these three quantitative measures, the study also proposes to demonstrate by way of qualitative data (e.g., teachers' self-reflective commentaries) that the professional development sequence into which this study has been inserted will provoke in at least some of its participating teachers an awareness of points at which they might improve their feedback practices.

Definitions of Terms

General-Impression Marking: Cooper (1977) describes *general impressing marking* as "the simplest" of holistic evaluative procedures, requiring "no detailed discussion of features and no summing of scores given to separate features" (p. 11). Instead, raters simply decide "where [each] paper fits within the range of papers produced for that assignment or occasion" (p. 12). In the case of this study, general-impression marks will be given according to the common practice of a percentage grade, whereby teachers will draw from their own professional experiences and their understanding of a simulated assignment context to describe the overall relative merits of each of a pair of essays.

Feedback: Adopting the conceptualization posited by Hattie & Timperley in their metastudy "The Power of Feedback" (2007), this project will define *feedback* as "information

provided by an agent (e.g., teacher, peer, parent, self, experience) regarding aspects of one's performance or understanding" (p. 81). Hattie & Timperley further illuminate their definition with a few meaningful introductory examples:

> A teacher or parent can provide corrective information, a peer can provide an alternative strategy, a book can provide information to clarify ideas, a parent can provide encouragement, and a learner can look up the answer to evaluate the correctness of a response. Feedback is thus a "consequence" of performance. (p. 81)

As stipulated by Brookhart (2008) based on her review of Hattie & Timperley and others (e.g., Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Butler & Winne, 1995; Kluger & DeNisi, 1996) the content of feedback includes *focus* (on the work itself, on the process used to complete the work, on the student's self-regulation, or on the student individually), *comparison* (to a criterion of success, to the norm of other students, or to the student's previous performances), *function* (description or evaluation), *valence* (positive or negative), *clarity* (clear or unclear to the student), and *specificity* (nitpicky, overly general, or just right), and *tone* (helpful and respectful or lecturing and bossy).

Limitations

Internal Validity: No single-instrument demonstration of an assessment procedure's implications for improved teacher feedback to students can disaggregate actual effects from effects that are merely artifacts of the study design, its instrument of data collection, the characteristics of the test subjects, and so on (Cook & Campbell, 1979). Moreover, although the study design strives to limit guessing within the treatment groups, and although the researcher is aware of the propensity of a researcher's expectations to skew perceived outcomes, the

contextual limitations of this study have place demands on the work that simply cannot be altered: its major data collection piece has been constrained by the limits of a 45-minute period on a teacher in-service day; it is not possible to administer the instrument to the two condition groups in separate locations; and the nature of the data set will be such that neither the researcher nor the trained assistant will be able to avoid noting which texts belong to which condition group.

That said, as a measure of how the teachers of the study's host district might alter their feedback in response to a procedural change in their grading practices, the study incorporates the potential for strong internal validity. In other words, we might learn a great deal about what *this district's* teachers are likely to do in similar feedback-and-grading situations within their own day-to-day work, for the subjects of the study are not teachers *similar to* the teachers of the site of interest; rather, they *are* the teachers of the site of interest. Within the limitations of the study, we actually could learn the probable response within this district to a revision of grading practices such as suggested by the study.

External Validity: It is likely that other secondary teachers will see mirrored in this study's focus several issues pertaining to their own grading and feedback practices. Across the United States, teachers simply do not have the time or energy to do "everything" that might be done each time student essays cross our desks. Therefore, it would not be surprising for readers to recognize and applaud the effort to consider how we might improve our practices so as to free up time from sorting-oriented activities so that we might apply more energy to learning-oriented ones. Along such a line, even where the study's data do not produce hypothesis-supporting results, this work might nevertheless inspire teachers and researchers to continue similar lines of thought, for no other reason than that they see it as a useful way of reframing ongoing

discussions about our assessment practices.

It is expected, however, that many teachers (even those within the study itself) may feel that placing more of the grading responsibility into the hands of students is likely to undercut teachers' sense of security that their grades are "accurate" or "fair." For this reason, the proposed study has included the data analysis for general-impression marking and interrater reliability. Yet if interrater reliability improves even at the cost of some grade inflation, the feedback practice introduced here may still have a certain degree of appeal to classroom teachers interested in students' perceptions of "fairness" (a student-oriented perspective that amounts to an informal impression of interrater reliability). Teachers would very likely see in the improved IRR a benefit worth the comparatively minimal effort to recalibrate their general-impression marking so that grades once again seem "true."

Positive Results: Positive results in this study, particularly with respect to the combination of H2 (IRR improves) and H3 (better feedback occurs), might indicate a reason for replicating this study's approach in other settings—including authentic, classroom-based settings—and with a wider variety of exemplar texts to see if its findings hold true across contexts. Additionally, should H1 (grades improve) prove true, a follow-up study might be warranted to show how the effects of better feedback could be sustained while implementing other procedures—such as periodic grader calibration—designed to avoid grade inflation.

Perhaps the most valuable hypothesis in this study is H3. Positive results here even in the absence of greater IRR and "loss of rigor" in grading might nevertheless warrant altered assessment practices, at least within the host district—which has recently re-cast is strategic plan in such a way as to place comparatively greater emphasis on whole-learner outcomes, less emphasis on grades and standardized assessments.

Negative Results: A false result in H3 may be an indicator of a faulty research design. Of particular concern is the potential for teachers not to take a simulated scoring activity—outside of their authentic work with actual students—seriously enough so as to provide feedback and scores that are reasonably faithful to their actual practices. It is also possible that the limited pool of research participants—approximately 30 middle school and 25 high school participants per condition—may not have sufficient power to illuminate actual differences under the experimental condition.

Alternatively, a false H3 might simply be a demonstration that the hypothesis is without merit. While it seems plausible that self-reflective student comments on a text would spur teachers toward providing richer feedback—speaking to the student and not simply to the text—it might be that that these student-provided will be ignored or even that they will increase teachers' frustration with the evaluative task.

Chapter Summary

Scarcity of time, low interrater reliability, and the challenge of providing feedback that is not simply sorting-oriented but also learning-oriented all work together to comprise a three-pronged problem for English language arts teachers every time they sit down to assess and provide comments to a stack of student papers. Drawing upon the research-based theoretical paradigm suggested by Hattie & Timperley (2007) and adopted by Brookhart (2008), the current study proposes to investigate whether under an experimental condition it may be possible to sufficiently free teachers from some of their sorting-oriented obligations so that they might engage in higher rates of providing feedback that supplements a focus on the student text with additional foci on the student's process of composing and revising that text and the student's self-regulation in the understanding and skill-development requisite to the writing task.

Moreover, the study proposes to investigate whether the number of comments that are positively framed and given in such a way as to be readable as "helpful" (respectful, preserving the student's agency as author) also increases under the experimental condition.

Such increases in the richness of teacher feedback may involve trade-offs. As teachers' attention may be distracted from sorting-oriented issues, they may tend to give higher general-impression scores, though perhaps with greater interrater reliability as measured by intraclass correlations.

Finally, the study proposes—by way of teachers' self-evaluative commentaries within the primary data set—to demonstrate the potential merits of a follow-up study, implementing in actual classrooms the assessment-and-feedback model simulated in the current project.

CHAPTER TWO: LITERATURE REVIEW

General-Impression Marking

General-impressing marking is a subset of essay evaluations known collectively as *holistic scoring*. Holistic scoring, much as the name implies, is an approach to evaluating essays that sets scorers' focus on a work's overall effectiveness, rather than constraining their reviews to specific analyses of individual traits within the writing (White, 1985). Scorers read a work and—usually guided by judgment-aiding rubrics and calibration to scoring norms (Diederich, 1974)—provide a single score representing the work's overall merit. When this work is done well, it is possible for holistic scoring methods to achieve interrater reliability levels greater than 0.8 (Diederich, 1974; Hillocks, 1986).

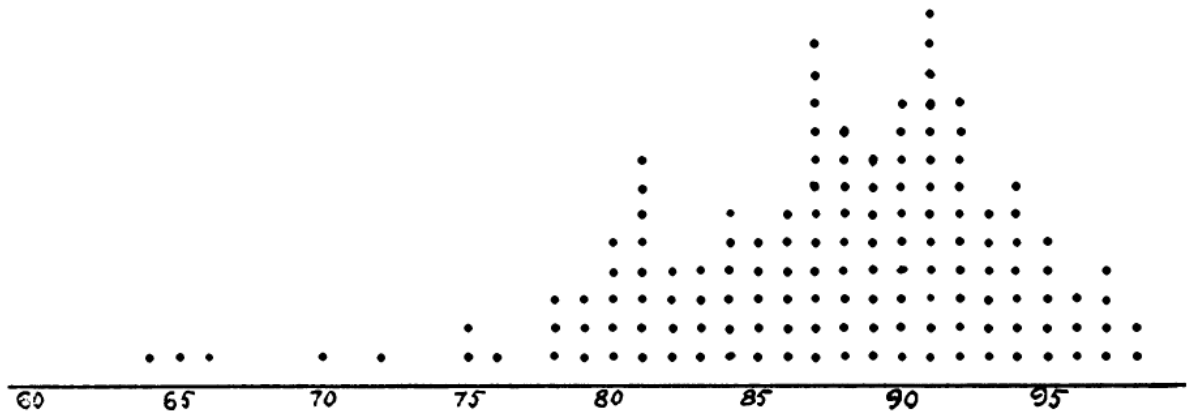
The practice, if not the fully articulated theory,[10] of holistic scoring extends back further than do reliability studies in English education. The earliest study consulted for this research project, Starch and Elliot's "Reliability of the Grading of High-School Work in English" (1912) makes use of an already familiar holistic scoring system in asking its study participants to evaluate two sample papers according to a 100-point grading scale, with a passing mark usually in the 70- to 75-point range. Results in the Starch and Elliot study, plotted along dot charts, demonstrate pictorially the outcomes of the scoring—in this case, the degree to which the scorers disagreed about the papers' overall merits—so that it is easy for readers to make on-

[10] In an interview with Bowman, Mahon, and Pogell (2004), writing assessment specialist Huot notes that although a fellow graduate student in the 80s claimed "holistic scoring was a practice without a theory" (p. 95), the theoretical foundations had been established at least by the 1960s with the publication of an ETS research bulletin by Godshalk, Swineford, and Coffman (1966). Huot goes on to note that the demonstration of these theoretical underpinnings fostered the transition from indirect tests of student writing in Advanced Placement exams.

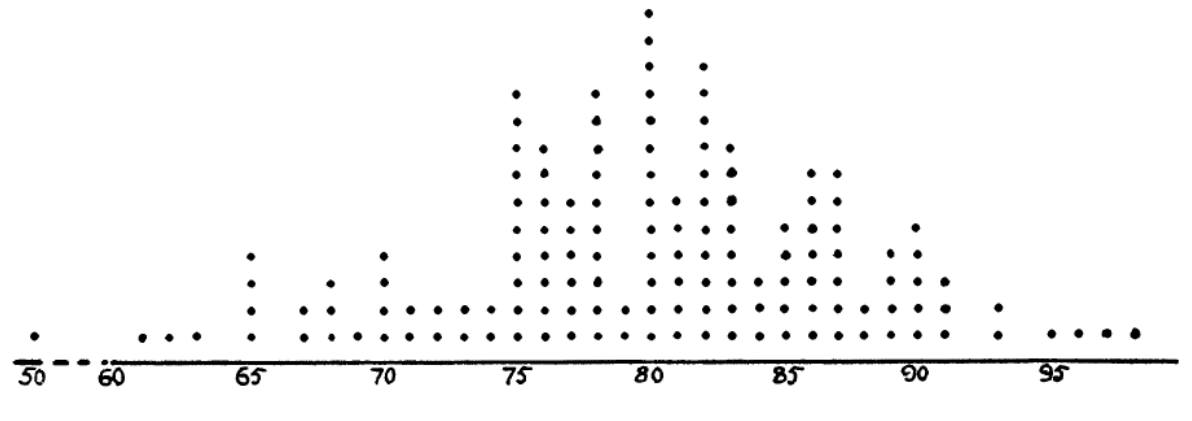
the-fly visual judgments about measures of central tendency and distribution, as seen in Figure 3. And in fact it is this sort of pictorial representation and the more precisely articulated statistical descriptions it supports that have made holistic scoring a frequently appearing characteristic of late twentieth-century, large-scale testing, such as the SAT, ACT, and AP Literature exams, as well as early versions of the NAEP writing exam.

Figure 2 Starch and Elliott's Distributions of Scores for Two High School Papers (1912)

Starch and Elliot, Figure 5 (1912, p. 451): Paper A. Passing grade 75. median 88.2. Marks assigned by schools whose passing grade is 70 are weighted by one point.



Starch and Elliott, Figure 6 (1912, p. 452): Paper B. Passing grade 75. Median 80.2. Marks assigned by schools whose passing grade is 70 are weighted by two points.



According to Applebee (1994), such large-scale direct assessments of writing became prominent in the 1970s, when on the grounds of "psychometric precision" (p. 41) they increasingly replaced indirect measures of writing ability such as multiple choice examinations of grammar and usage. In the years since, we have learned that the claims for greater precision were unfounded,[11] but the pedagogical outcomes of this paradigm shift in testing have been enormous. Applebee notes that dependence on indirect measures of writing assessment "amount[ed] to a decision to emphasize the teaching of word, sentence, and paragraph skills, rather than to emphasize purposeful thinking and writing" (p. 41). Applebee further argues that

[11] On this, see also Huot's comments in Bowman, Mahon, and Pogell (2004), in which he discusses the reliability of indirect-methods exams, the challenges of interrater reliability for direct assessments, the "perfect scoring reliability" of computer scoring programs (p. 96), and the recent assessment research shifts away from reliability to the validity of various assessment approaches. In these shifts, questions of interrater reliability find themselves couched in the specific, real-world context s of the assessments themselves. When such contexts are allowed into consideration, individual readers' professional backgrounds and academic commitments are brought to the foreground and used purposefully rather than being subject to generic controls against unwanted variance.

For instance, if a university program designs an assessment for placing incoming students into a particular course—such as a regular course in first-year composition course or a basic writing skills remedial course—various exam scorers will have predictably classifiable and useful reactions to students' writing samples, based not only on their normed and/or idiosyncratic responses to the text itself, but also on their familiarities with the course in question and their experience-informed understandings of the academic and practical outcomes of a vote toward placing the student in one course or the other. According to Huot (Bowman, Mahon, & Pogell, 2004; also, Huot, 1996), such an assessment approach has much to argue for its validity in that it honors the specifically contextualized world for which the assessment has been prepared, from which the scorers make their judgments, and in which the students will subsequently continue their academic learning. This sort of contextualized approach to writing assessment seems worlds away from using an ACT or SAT sub-score for course placement.

teaching and learning have improved dramatically as a result of our professional community having chosen a new assessment paradigm: "Twenty years ago," he writes, "one could teach writing without asking students to write. Due in part to changes in the format of writing tests, that is no longer true today" (p. 41).

Yet despite the very encouraging curricular outcome resulting from the shift toward direct measures of writing proficiency, other English education specialists identified problems holistic approaches almost immediately. For one, direct writing measures like the NAEP were criticized for the ways they valorized certain written products over others, as well as the conditions under which these products were produced. "As teachers have embraced new approaches to writing instruction," admits Applebee, "NAEP writing tasks (typically taking fifteen minutes to an hour, on a set topic, with little room for planning or revision) have become at best an imperfect reflection of curricular wisdom" (p. 41). Although Applebee notes that from a psychometric point of view the problem is irrelevant, "because there is no evidence to suggest that students' performance *relative to that of other students* changes" under more authentic testing conditions, he concedes that there may still be a reasonable argument against traditional holistic measures' continued emphasis in American educational systems. As it turns out, this argument is a close cousin to the one several decades ago that spurred us to transform our testing focus from multiple-choice items to written passages:

> The better argument against current approaches to assessment is on curricular rather than psychometric grounds. If there is an emerging consensus about the value of writing assessments in which students have time to engage thoughtfully in planning and revision activities, then that consensus *must* be reflected in the ways which student performance is assessed. For if assessment remains out of

> alignment with curriculum, it is curriculum, not assessment, that will suffer. (p. 42)

Put another way, timed, non-revised, frequently "academic," holistic writing assessments, though they may represent an improvement over indirect methods of gaining insights about student abilities, are "based upon a set of assumptions and beliefs irrelevant to written communication," and they may need to be replaced by "assessment theories and practices which are consonant with our teaching and research" (Huot, 1996, p. 564).

It is for this reason among others that various alternative measures have gained popularity over the last several decades. Huot (1996) notes that these measures generally break down into two basic categories. First are the alternative measures specifically designed to serve as *placement exams*—for which generically generated texts judged by interchangeable, trained, calibrated readers, have sometimes been replaced by task-specific writings reviewed by raters whose "most immediate and extensive teaching experience" (Huot, 1996, p. 553) in the course of interest makes them ideal judges with respect to the likely outcomes of a particular placement decision. Sometimes, these contextually driven judgments have been made according to a two-tier system, in which a first reader determines whether placement in an introductory course is advisable and then, if not, subsequent readers determine which course placement makes the most sense. In each assessment variation, according to Huot, "these contextualized forms of placement assessment are sound because teachers make placement decisions based upon what they know about writing and the curriculum of the courses they teach. Placement of students in various levels of composition instruction is primarily a teaching decision" (p. 554).

A second group of alternative writing measures are those functioning as *exit exams* and *program assessments*. Portfolio reviews usually fall into this category, sometimes being used to

make decisions about students, as when educators must "determine whether or not students should move from one course to another" (Huot, 1996, p. 554; Durst, Roemer, and Schultz, 1994) or when they must demonstrate that students have satisfied their overall program completion requirements, as in teacher education programs (Zeichner & Wray, 2001). At other times, portfolio assessments provide data for review teams to make judgments about programs themselves (Allen, 1995; Huot 1996). Portfolio-based assessments have admirers in that they "offer an opportunity to examine classroom-based samples of literature behavior in reading and writing, chosen by the student and teacher to represent a broader spectrum of performance than can ever be sample in an examination situation" (Applebee, 1994, p. 44).[12] But Applebee suggests that portfolios are not without their problems, among which are the challenges of disaggregating individual from group performances, of determining which pieces belong in the portfolio and which are to be excluded, of determining who makes the inclusion decision, of how the portfolio is to be evaluated, and the degree to which they allow for meaningful longitudinal or cross-sectional comparisons. From a psychometrician's point of view, Applebee concludes, "portfolios are not at the moment very popular" (p. 45).

An altogether different sort of alternative assessment suggested for use simultaneously as an exit piece and for program review is the *generative prompt*, which asks for students to respond to and analyze their own experiences throughout a course of study (Condon, 2009). In

[12] Supovitz and Brennan (1997) note that portfolio-based assessment—although having a "mixed effect on equalizing the differences in performance of students with different backgrounds and experiences"—are promising on the grounds of "focusing instruction on higher-order thinking skills, providing useful feedback to teachers about student thought processes, and emphasizing real-world skills and problem-solving" (p. 498).

that such an assessment style may ask students to provide information about the courses they considered most influential to their own learning, the best teaching strategies and assignments they encountered, their most memorable instructors, or the degree to which overarching program goals have been integrated at the course level, Condon argues that it has the potential to make "the assessment enterprise important to the institution in ways that move beyond the need for data about students' learning experiences"; it is, he concludes, an assessment approach that is in "our enlightened self-interest" (p. 153).

To the degree that enlightened self-interest becomes the force that guides us in the direction of better assessments of writing—both at the classroom and program levels—there are in fact many well documented recommendations to which we should adhere. Our assessments should, whenever, possible, involve:

- Tasks that spring from local curricula and classrooms rather than being imposed upon them from the outside (Applebee, 1994; Bowman, Mahon, & Pogell, 2004; CCCC Committee on Assessment, 1995; Huot, 1996); despite the potential for stakeholder biases, Barlow, Liparulo, and Reynolds (2007) advocate that all stakeholder-participants in an assessment program be part of the design and implementation processes so that everyone can share "full confidence in the process and results" (p. 52), thus minimizing the sense that writing assessment is somebody else's responsibility, that its outcomes and next steps toward remedies somebody else's problems; Huot notes that such an approach is "a lot cheaper than conventional writing assessments because you don't have to pull anchor papers, you don't have to create rubrics, you don't have to norm people, you don't have to renorm people, and then when you get scores, you don't have to sum the

scores or do split resolvers or set cut scores or place people based on the scores. You have teachers reading student writing and then making the decision directly" (Bowman, Mahon, and Pogell, 2004).[13]

- Tasks involving "higher literacy," prioritizing rich, interpretive thinking and the construction of well-defended points of view (Applebee, 1994, p. 45; Huot, 1996).
- Tasks including opportunities for reconsideration and revision of drafts (Applebee, 1994), even going so far as to leverage students' individual abilities through meaningful social interactions such as those that occur during the discussion and feedback surround their written texts (CCCC Committee on Assessment, 1995).
- Multiple collection opportunities across a wide range of tasks (CCCC Committee on Assessment, 1995; Applebee, 1994; Barlow, Liparulo, & Reynolds, 2007).
- Tasks designed in such ways as to avoid misrepresenting the skills of students from marginalized groups (CCCC Committee on Assessment, 1995; Condon, 2009; see review of studies on testing gaps in Hillocks, 2006).
- Judgments that spring from expert teachers' beliefs about reasonable expectations from students, not just those of a well-intentioned but under-informed policy makers (Applebee, 1994, pp. 44, 46); in this, our assessments must focus on

[13] For a brief alternative statement on the economic pressures that guide standardized testing—which "virtually assure that the lowest form of assessment that provides the appearance of thoroughness and the greatest economy will prevail"—see Condon's introductory comments in "Looking Beyond Judging and Ranking: Writing Assessment as a Generative Practice" (2009, p. 142).

> producing valid and reliable results within the contexts they are actually to be used for making changes to teaching and learning (Barlow, Liparulo, & Reynolds, 2007; Huot, 1996); as Huot reminds us, "[t]he people who are best qualified to decide who belongs in [particular] courses are the people who teach those courses" (Bowman, Mahon, & Pogell, 2004).

Most importantly, as teachers, administrators, and assessment development teams design written measures of writing proficiencies, our first and primary consideration must always be the attempt to understand what influences—intended and otherwise—the assessment itself will have on curriculum and instruction. Our assessments, in other words, need "systemic validity" (CCCC Committee on Assessment, 1995). We need no more tests that stultify the curriculum.

Although the more recent reconsiderations of holistic writing assessment have proceeded primarily according to matters of validity, earlier suggestions responded to the lack of information provided in the holistic scores themselves, which are good for sorting and ranking but which don't tell us much about the constituent elements of any student's work. Along this line of critique, Lloyd-Jones (1977) suggested *primary-trait scoring* on the merits of its ability to highlight features of writing—"the separate elements, devices, and mechanisms of language" (p. 33)—that are lost in the totalizing approach of holistic scoring.[14] Figure 3 provides an example

[14] Following the academic discourse by which holistic scoring has been defined against other models is somewhat challenging. Both Cooper (1977) and Lloyd-Jones (1977), for instance place primary-trait scoring under the classification of holistic models. Applebee (1994), however, speaks of primary-trait assessment as "radically different from that of general-impression or holistic scoring" (p. 43). At this stage in my own learning, I'm inclined not only to land with Applebee but also to wonder of others among Cooper's so-called holistic measures are more properly considered analytic. While *general-impression marking, essay scale*—"a series of complete pieces

of three primary traits from a scoring guide in Lloyd-Jones' introduction to the approach.

Figure 3 Examples from a Primary-Trait Scoring Guide (Lloyd-Jones, 1977, pp. 52-53)

***Directions for Student Writers*:** Look carefully at the picture. These kids are having fun jumping on the overturned boat. Imagine you are one of the children in the picture. Or if you wish, imagine that you are someone standing nearby watching the children. Tell what is going on as he or she would tell it. Write as if you were telling this to a good friend, in a way that expresses strong feelings. Help your friend FEEL the experience too. Space is provided on the next three pages.

Use of Dialogue	
0	Does not use dialogue in the story.
1	Direct quote from one person in the story. The one person may talk more than once. When in doubt whether two statements are made by the same person or different people, code 1. A direct quote of a thought also counts. Can be in hypothetical tense.
2	Direct quote from two or more persons in the story
Point of View	
0	Point of view cannot be determined, or does not control point of view.
1	Point of view is consistently one of the five children. Include "If I were one of the children . . . " and recalling participation as one of the children.
2	Point of view is consistently one of an observer. When an observer joins the children in the play, the point of view is still "2" because the observer makes a sixth person playing. Include papers with minimal evidence when difficult to tell which point of view is being taken.

arranged according to quality" (p. 4)—and Elbow's (1973) *center of gravity response* are clearly totalizing approaches to essay assessment, *primary trait scoring*, *analytic scale*, and *dichotomous scale* all seem to share the characteristic of bypassing a whole-piece judgment in favor of drawing the reader's attention to specific traits, treating these as subscales and then perhaps summing them for a "holistically" sortable value.

If definitions of *holistic* have shifted in the last three decades, perhaps the shift involves what Lloyd-Jones (1977) refers to as *atomistic* methods of writing assessment, which seem to include tests of vocabulary, usage, and syntax rather than of writing, per se—what we would call today indirect measures of writing. It would seem that in the 1970s *holistic* was taken to mean any sort of assessment that involved "relative pervasive elements of discourse (concreteness, coherence, liveliness), which must be described by trained human readers" (Lloyd-Jones, 1977, p. 36), whether that assessment was given according to general impression or some sort of composite of subscales.

Tense	
0	Cannot determine time, or does not control tense. (One wrong tense places the paper in this cateogry, excepted drowned in the present.)
1	Present tense—past tense may also be present if not part of the "main line" of the story.
2	Past tense—If a past tense descirption is acceptable brought up to present, code as "past." Sometimes the present is used ot create a frame for past events. Code this as past, sine the actual description is in the past.
3	Hypothetical time—Papers written entirely in the "If I were on the boat" or "If I were there, I would." These papers often include future references such as "when I get on the boat I will." If part is hypothetical and rest past or present and tense is controlled, code present or past. If the introduction, up to two sentences, is only part in past or present then code hypothetical.

Applebee (1994) describes primary-trait scoring in contrast to holistic scoring not only in that it "captures an aspect of performance different from general-impression marking" (p. 43) but also because it operates from a different theoretical starting point with respect to assessment; for whereas holistic scoring assumes a normal distribution of scores, primary-trait assessments begin from the position that it is "quite possible that *no one* in a given population will be able to complete a particular task successfully, while on other tasks *everyone* may be successful" (p. 43).

Like primary-trait scoring, *analytic* scoring "breaks into" a text to view it by component elements rather than as a totality. Cooper (1977) points to Diederich (1974) for an example of analytic scoring derived from his study of the factors pertaining to ratings of writing ability, as seen in Figure 4. Diederich offers these as a "checklist to improve the consistency of [teachers'] ratings," but does so with a caution: "I have never had much confidence in any scheme for rating papers that does not involve comparison with independent ratings of another person and discussion of papers on which there is a substantial difference of opinion" (p. 53). In other words, it is not the ratings instrument itself that makes for reliable ratings, but shared sets of commitments, understandings, and experiences among evaluators that are likely to generate better agreement in scores.

Figure 4 An Analytic Scoring Guide (Diederich, 1974, p. 54)

Teacher Rating Slip for Student Essays: Note the double weighting for ideas and organization, on account of these qualities' emphasis in the courses for which the essays were written

Topic_____ Reader_____ Paper_____

	Low		Middle		High	
Ideas	2	4	6	8	10	
Organization	2	4	6	8	10	
Wording	1	2	3	4	5	
Flavor	1	2	3	4	5	_____
Usage	1	2	3	4	5	
Punctuation	1	2	3	4	5	
Spelling	1	2	3	4	5	
Handwriting	1	2	3	4	5	_____
					Sum	_____

Cohen's dichotomous scoring (1973), too, analyzes individual traits in a written work, but instead of giving numerical quality ratings, this approach merely provides "yes" and "no" characterizations of whether the desired trait is present, as in Figure 5. Cooper (1977) notes that although dichotomous scales might not be discriminating enough to provide reliable results for individual writers, they would be "quite promising" (p. 9) for judging the overall success of groups, as in the case of making judgments about program effectiveness. Cohen himself reaches similar conclusions, finding particular value in how the development of this approach—as a research project embedded within its host institution's writing program review—not only has the ability to make judgments about program effectiveness but also to provide what amounts to instruction-transforming professional development for its participating instructors.

Figure 5 A Dichotomous Scoring Guide (Cohen, 1973, p. 359)

Teacher Rating Slip for Student Essays: Note the double weighting for ideas and organization, on account of these qualities' emphasis in the courses for which the essays were written

SCORE SHEET

	YES	NO	
Content I.	——	——	1. Ideas themselves are insightful.
	——	——	2. Ideas are creative or original.
	——	——	3. Ideas are rational or logical.
	——	——	4. Ideas are expressed with clarity.
Organization II.	——	——	5. There is a thesis.
	——	——	6. Order of thesis idea is followed throughout the essay.
	——	——	7. Thesis is adequately developed.
	——	——	8. Every paragraph is relevant to the thesis.
	——	——	9. Each paragraph has a controlling idea.
	——	——	10. Each paragraph is developed with relevant and concrete details.
	——	——	11. The details that are included are well ordered.
Mechanics III.	——	——	12. There are many misspellings.
	——	——	13. There are serious punctuation errors.
	——	——	14. Punctuation errors are excessive.
	——	——	15. There are errors in use of verbs.
	——	——	16. There are errors in use of pronouns.
	——	——	17. There are errors in use of modifiers.
	——	——	18. There are distracting errors in word usage.
	——	——	19. The sentences are awkward.

Primary-trait, analytic, and dichotomous scoring, then, function within one class of responses to holistic writing—each of them assuming the holistic assessment's written task and depersonalized scoring as constants, but fracturing the "whole essay" outcomes of student

writing into more finely detailed sets of characteristics so as to provide instructionally useful feedback to writers and their teachers. The other class of responses—including specifically contextualized pieces, generative essays, and portfolio assessments, among others—involves approaches in in which assessment designers closely align their tasks to current beliefs about the teaching and learning of writing. And just as is the case with the "earlier" revisions to holistic scoring, these approaches provide much richer information than mere ranking and sorting—allowing not only students and teachers, but even program design teams access to any number of insights that can extend (everyone's) learning beyond the point of assessment itself.

These practical and philosophical challenges notwithstanding, holistic scoring continues to enjoy a prominent place in local and large-scale testing on account of its ability to demonstrate a distribution of scores. Its job, in other words, is to help us sort and compare outcomes. Frequently, the work of outcomes-sorting is directed at students, as when teachers make judgments about papers that translate into percentage scores in a grade book, which can then combined with hundreds of similar scores across four years of high school or college to determine eligibility for accolades like the National Honor Society, valedictorian, summa cum laude, Phi Beta Kappa, and so on. At other times, administrators use pooled holistic scores, too, perhaps when they compare the outcomes of student cohorts to determine whether individual teachers are effectively preparing students for standardized exams like the Kansas Reading Assessment. At least one teacher in my building is *sure* that somebody in district office is just waiting to land hard on her if her AP scores drop below their historically high levels. Researchers, too, find holistic scores a useful tool. Like classroom practitioners and supervisors we sometimes focus on student differences, but sometimes we're also interested in the outcomes of the sorters themselves, as in studies like those conducted by Starch and Elliot (1912),

Diederich (1974), or Chase (1983). Such is the case of this study, too, where a simple, holistic score—a judgment about the earned percentage of whatever might hypothetically be "full credit" on the assignment in question—might help us understand if teachers value students work at a higher or lower level simply because the student has annotated a for-submission essay with additional, personally relevant self-evaluative commentaries.

It is for these reasons that this study has adopted the use of a certain class of holistic scoring—*general-impression marking*—as a variable for detecting whether the feedback-providing routine at the heart of its study has any secondary effects on teachers' impressions about the overall merits of the essays to which they are responding. Cooper (1977) describes *general impressing marking* as "the simplest" of holistic evaluative procedures, requiring "no detailed discussion of features and no summing of scores given to separate features" (p. 11). Instead, raters simply decide "where [each] paper fits within the range of papers produced for that assignment or occasion" (p. 12). In this study, general-impression marks will be given according to the common practice of a percentage grade, whereby teachers will draw from their own professional experiences and their understanding of a simulated assignment context to describe the overall relative merits of each of a pair of essays.

Feedback

Feedback refers to "actions taken by (an) external agent(s) to provide information regarding some aspect(s) of one's task performance" (Kluger & DeNisi, 1996), and it is "among the most critical influences on student learning" (Hattie & Timperley, 2007). For the purposes of this study, *feedback* will be defined according to Hattie & Timperley's conceptualization as "information provided by an agent (e.g., teacher, peer, parent, self, experience) regarding aspects of one's performance or understanding" (p. 81). As stipulated by Brookhart (2008) based on her

review of Hattie & Timperley and others (e.g., Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Butler & Winne, 1995; Kluger & DeNisi, 1996) the content of feedback includes *focus* (on the work itself, on the process used to complete the work, on the student's self-regulation, or on the student individually), *comparison* (to a criterion of success, to the norm of other students, or to the student's previous performances), *function* (description or evaluation), *valence* (positive or negative), *clarity* (clear or unclear to the student), and *specificity* (nitpicky, overly general, or just right), and *tone* (helpful and respectful or lecturing and bossy). This study will be using Brookhart's stipulation to consider the feedback provided in teachers' written responses to student essays.

Kluger and DeNisi (1996) offer what is perhaps the best—though not easiest—introduction to the understanding that not all feedback is good feedback. Entering an already decades-old conversation, they found a field nearly saturated with the mistakenly overgeneralized notion that a student's *knowledge of performance* increases learning and motivation (Ammons, 1956). Examining this claim closely, Kluger and DeNisi found not only inconsistencies within the originating author's own evidence, but in the scholarly tradition flowing from it. Citing various sources composed through at least the late 1980s (e.g., Ashford & Cummings, 1983; Harris & Rosenthal, 1985; Pritchard, Jones, Roth, Stuebing, & Ekeberg, 1988), Kluger and DeNisi found that "scholars continue to ignore findings suggesting that FI [i.e., *feedback intervention*] effects on performance are highly variable" (p. 256). In response, they developed a meta-analysis "to determine whether the variance [in average FI effects on performance] merely reflects sampling-error variance . . . or some *true* negative effects of FIs on performance" (p. 257). Their findings demonstrated that the average effect size of feedback was indeed positive, and moderately high (.41), even with the downward drag of more than one-third

of the interventions, including discouragement (-0.14), praise (0.09), and feedback given following the completion of comparatively complex tasks (0.03). Among the strongest positive influences on feedback effects were those involving velocity—i.e., feedback describing changes from previous attempts—(0.55) and the setting of goals (0.51), those drawing attention to correct performances (.43), and those given following the completion of memory-oriented tasks (0.69) as opposed to physical (-0.11), or rule-following tasks (0.19).

Although Kluger and DeNisi's study provides a meaningful articulation of productive versus harmful feedback interventions, its usefulness for classroom teachers is limited, couched as their article is in the language of a particularly demanding meta-analysis surveying 131 feedback studies across a variety of fields. Fortunately, other studies have followed, including a somewhat more accessible work by Hattie and Timperley (2007). Hattie and Timperley open their discussion with a report on Hattie's (1999) review of 357 meta-analyses of interventions affecting learning outcomes in schools. This synthesis demonstrated the average effect of schooling to be 0.40 (SE = 0.05) while that the average effect of feedback itself was 0.79, or twice that of school in general. Having drawn attention to the general efficacy of feedback, the authors briefly revisit Kluger and DeNisi's (1996) findings, providing the following clarifying summary of their findings:

> Across all comparisons, it appears that the power of feedback is influenced by the direction of the feedback relative to performance on a task. Specifically, feedback is more effective when it provides information on correct rather than incorrect responses and when it builds on changes from previous trails. The impact of feedback is also influenced by the difficulty of goals and tasks. It appears to have the most impact when goals are specific and challenging but task complexity is

> low. Praise for task performance appears to be ineffective, which is hardly surprising because it contains such little learning-related information. It appears to be more effective when there are low rather than high levels of threat to self-esteem, presumably because low-threat conditions allow attention to be paid to the feedback. (Hattie & Timperley, 2007, pp. 85-86).

With both this review of Kluger and DeNisi and also Hattie's findings about the efficacy of feedback in general, Hattie and Timperley turn to "identifying the conditions that maximize the positive effects on learning" (p. 86).

Perhaps most important of these conditions is actually an awareness of what feedback is supposed to do. "The main purpose of feedback," write Hattie & Timperley (2007), "is to reduce discrepancies between current understandings and performance and a goal" (p. 86). In this, feedback functions as one of several useful strategies available to students for bridging the gaps between their current abilities and their desired outcomes. Some of these strategies originate from students themselves. For example, they can apply increased effort, "particularly when the effort leads to tackling more challenging tasks or appreciating higher quality experiences rather than just doing 'more'" (p. 86). Other student-initiated strategies are improved error-spotting abilities, and better problem-solving and task-completion strategies.

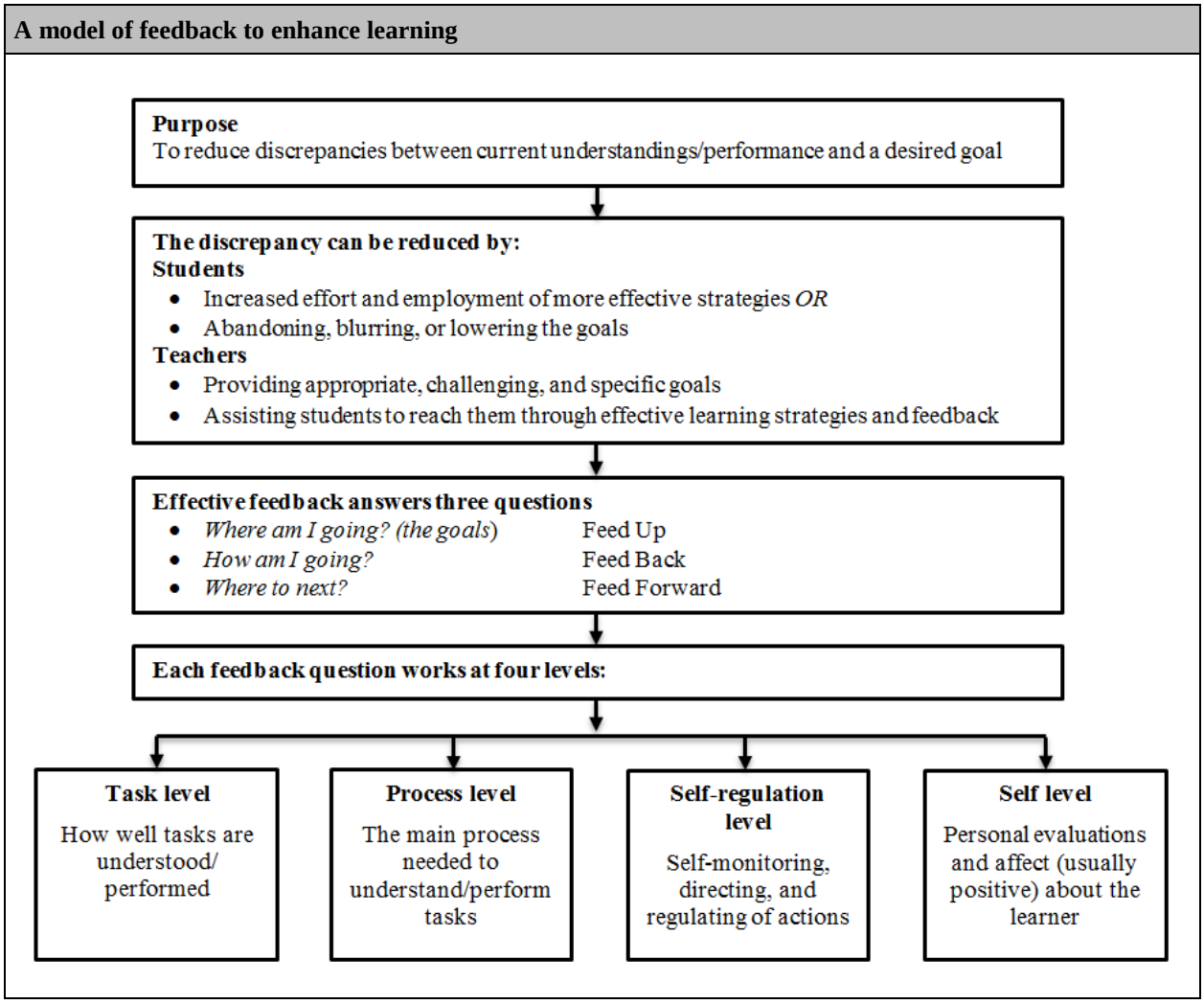
Teachers, too, have access to strategies that can help students bridge the gaps between their current achievements and future goals. First, teachers can ensure that the goals set for their students are "appropriate, challenging, and specific" (p. 87; see also Kluger & DeNisi, 1996; Lock & Latham, 1984). Second, "teachers can also assist by clarifying goals, enhancing commitment or increased effort to reaching them through feedback" (p. 87). Such feedback can help students discard their less-useful hypotheses about what is causing their gaps (Sweller,

1990); it can also assist in the development of better self-regulation and error spotting (Hattie, Biggs, & Purdie, 1996).

To accomplish the purpose of reducing discrepancies, feedback addresses three central questions: "*Where am I going? (What are the goals?), How am I going? (What progress is being made toward the goal?),* and *Where to next? (What activities need to be undertaken to make better progress?)*" (Hattie & Timperley, 2007, p. 86). Further, it addresses these questions among four levels of consideration—*task performance*, *understanding how to do a task*, *regulatory or metacognitive processes*, and the *personal* level. When providing feedback, teachers must ensure that the feedback operates in academically useful ways. This means, in part, that their feedback should have a sort of inherently dynamic character. Ideally, "feedback aimed to move students from task to processing and then from processing to regulation is most effective" (p. 91). When feedback, in fact, does not facilitate such movement from task to regulation, long-term outcomes suffer. Additionally, while effect feedback must direct students' attention to the work, the process, or to self-regulation, it should not focus on the student as an individual. Feedback about the individual—e.g., "Good girl!" "How smart you are!"—is ineffective "because it carries little information that provides answers to any of the three questions and too often deflects attention from the task" (p. 96). Taken all together, this multi-component rationale for and approach to feedback comprises Hattie & Timperley's model for feedback, as represented in Figure 6.

But what teachers need if they are to be providers of effective feedback, however, is something much more concretely accessible than is provided in Hattie & Timperley's framework. Like the students they serve, teachers need goals that are clear and understandable to non-specialists. They also need examples of what good and bad feedback look like. Such clarity

Figure 6 A Model for Feedback (Hattie & Timperley, 2007, p. 87)



and such examples are what makes Brookhart's *How to Give Effective Feedback to Your Students* (2008) an effective model for use in the current study. Brookhart's discussion of feedback extrapolates from insightful readings Kluger and DeNisi (1996), Hattie and Timperley (2007), and others to help classroom teachers bridge their own gaps between current performance and desired goals for composing instructional feedback. And while Brookhart's discussion doesn't itself constitute "research," it nevertheless adheres to research findings in way

that is at once both academically meaningful for this study, and also approachable as a resource for the English language arts teachers who have participated in this study's data collection and professional development piece. It is worth noting, furthermore, that Brookhart is an accomplished researcher in her own right (Brookhart, 2001; Brookhart & Devoge, 1999; Brookhart & Freeman, 1992), who can be trusted to "popularize" others' work with reasonable fidelity to their actual research findings.

Brookhart opens with a recognition of feedback's "Jekyll-and-Hyde character" in that not all feedback is helpful and that "because students' feelings of control and self-efficacy are involved, even well-intentioned feedback can be very destructive" (p. 2). She then quickly surveys research from Kluger and DeNisi (1996), Hattie and Timperley (2007), and others. Having set this context, Brookhart moves on to the core propositions of her work, that four *strategies* and seven *content characteristics* pertain to the delivery and content of feedback. The four strategies of feedback are *timing*, *amount*, *mode*, and *audience*, as demonstrated in Figure 7. While the four strategies are no doubt important, this study will focus on the following seven content characteristics of feedback, also demonstrated in Figure 8:

- *Focus*: Ideal feedback is focused either on the work itself, on the student's process in completing that work, or on the student's ability to self-regulate with respect to the gaps between current and "ideal" writing outcomes. It is, moreover, "suited to the draft we are reading" (Sommers, 1982, p. 155; see also Horvath, 1984), providing both the scope and detail that are most fit at the student's current achievement with the draft, while also maintaining relevance to the established writing values of the course (Dohrer, 1991). Feedback should avoid focusing on the student, individually, whenever possible.

Figure 7 Feedback Strategies (Brookhart, 2008, p. 5)

Feedback Strategies Can Vary In . . .	In These Ways . . .	Recommendations for Good Feedback
Timing	• When given • How given	• Provide immediate feedback for knowledge or facts (right/wrong) • Delay feedback slightly for more comprehensive review of student thinking and processing. • Never delay feedback beyond when it would make a difference to students. • Provide feedback as often as is practical, for all major assignments
Amount	• How many points made • How much about each point	• Prioritize—pick the most important points. • Choose points that relate to major learning goals. • Consider the student's developmental level.
Mode	• Oral • Written • Visual/demonstration	• Select the best mode for the message. Would a comment in passing the student's desk suffice? Is a conference needed? • Interactive feedback (talking with the student) is best when possible. • Give written feedback on written work or on assignment cover sheets. • Use demonstration if "how to do something" is an issue or if the student needs an example.
Audience	• individual • Group/class	• individual feedback says, "The teacher values my learning." • Group/class feedback works if most of the class missed the same concept on an assignment, which presents an opportunity for reteaching.

- *Comparison:* Feedback should provide student writers with points of comparisons between either the current draft and an established criterion, or between the current draft and previous or imaginable successive drafts. In this multiple-draft view is the implicit requirement that learning-oriented feedback be given within the context of courses allowing multiple rewrites for better outcomes (Burkland and Grimm, 1986; Freedman, 1987; Bardine, Bardine, & Deegan, 2000). It is not advisable for feedback to make comparisons among students, as this may

reinforce a sense of "winners" and "losers" in the educational process.

- *Description not Evaluation*: Good feedback is primarily descriptive in nature, minimizing evaluative comments. To achieve this goal, such feedback needs to avoid being readily interpretable as given to justify grades (Dohrer, 1991; Elbow, 1997), perhaps by being given in the absence of grades altogether, as is the case in formative assessments (Horvath, 1984; Burkland & Grimm, 1986). It needs, as well, to be readable as factual rather than opinionated in nature (Lynch & Klemans, 1978).
- *Clarity:* Comments must be expressed in language that non-expert writers can decode (Dohrer, 1991; Lynch & Klemans, 1978; Sommers, 1982; Land & Evans, 1987; Straub, 1997).
- *Specificity*: Additionally, comments must express specifically where successes and challenges lie and perhaps offers specific guidance as to next steps in the revisions process (Lynch & Klemans, 1978; Sommers, 1982; Land & Evans, 1987; Straub, 1997).
- *Helpfulness in Tone* (Bardine, Bardine, & Deegan, 2000; Lynch & Klemans, 1978): Successful feedback avoids over-emphasizing what students are doing wrong but looks as well for opportunities to illuminate show successes (Burkland and Grimm, 1986; Dragga, 1988; Daiker, 1989; Gee, 1972; Straub, 1997)
- *Positive Valence*: Good feedback is "positive" even when delivering bad news about current outcomes, not simply pointing out error but actually going so far as to offer suggestions for improvement (Straub, 1997). Perhaps these comments are delivered in such a way that helps students prioritize what is of greater or lesser

> importance for the next round of revisions (Sommers, 1982; Fuller, 1987); yet in doing so, they work hard to preserve students' agency and to avoid imposing teachers' purposes on students' writing (Sommers, 1982; Burkland and Grimm, 1986; Straub, 1997).

In that Brookhart's retransmission of earlier studies provides teachers with a clearly expressed set of feedback characteristics, as well concrete examples of what useful feedback actually looks like, I have chosen to use its language and framework as the basis for making qualitative judgments about feedback in this study.

Yet despite the fact that optimal feedback exists within a playing field of research-proven parameters and that it is efficacious in numerically demonstrable ways, we probably wouldn't be serving students well merely to proclaim to their teachers as if from on high, "Go, thou, and do likewise." As with many top-down mandates in education, achieving best-practices feedback by fiat is likely to be a low-percentage approach. A major reason for this likelihood involves what teachers know already about feedback and what additional knowledge they can add into their repertoires. On one hand, practitioners' pre-service training has very likely left them too inexperienced with research on feedback and its outcomes for them to incorporate any but the barest lessons from even a well-designed professional development session. On the other hand, teachers' time is so overburdened with things-to-be-done that they are quite unlikely to enjoy the time and energy to reflect thoroughly on their current practices, investigate alternatives landing closer to "best practices," and to try these out with our own students.

But perhaps the greatest complication to a system-wide improvement of feedback is that teachers don't simply "do" feedback in an impersonal, rather mechanical sort of way that responds well to top-down initiatives or one-off professional development sessions. Offering

Figure 8 Feedback Content (Brookhart, 2008, pp. 6-7)

Feedback Content Can Vary In . . .	In These Ways . . .	Recommendations for Good Feedback
Focus	• On the work itself • On the process the student used to do the work • On the student's self-regulation • On the student personally	• When possible, describe both the work and the process—and their relationship. • Comment on the student's self-regulation if the comment will foster self-efficacy. • Avoid personal comments.
Comparison	• To criteria for good work (criterion-referenced) • To other students (norm-referenced) • To student's own past performance (self-referenced)	• Use criterion-referenced feedback for giving information about the work itself. • Use norm-referenced feedback for giving information about student processes or effort. • Use self-referenced feedback for unsuccessful learners who need to see the progress they are making, not how far they are from the goal.
Function	• Description • Evaluation/judgment	• Describe. • Don't judge.
Valence	• Positive • Negative	• Use positive comments that describe *what* is well done. • Accompany negative descriptions of the work with positive suggestions for its improvement.
Clarity	• Clear to the student • Unclear	• Use vocabulary and concepts the student will understand. • Tailor the amount and content of feedback to the student's developmental level.
Specificity	• Nitpicky • Just right • Overly general	• Tailor the degree of specificity to the student and the task. • Make feedback specific enough so that students will know what to do but not so specific that it's done for them. • Identify errors or types of errors, but avoid correcting every one (e.g., copyediting or supplying right answers), which doesn't leave students anything to do.
Tone	• Implications • What the student will "hear"	• Choose words that communicate respect for the student and the work. • Choose words that position the student as the agent. • Choose words that cause students to think or wonder.

effective feedback is quite unlike the work of learning to use a voicemail system, proctoring a standardized exam, taking attendance, monitoring the halls during passing periods, or even checking the correctness of multiple-choice and short answer responses on a quiz. All of these are rather flat, routinizable tasks for which it might be meaningful to draw clear lines in the sand about such matters as promptness, accuracy, and time on task. But feedback is different. For although the strategies of feedback involve easily stipulated parameters—*timing, amount, mode,* and *audience* (Brookhart, 2008)—the communicative task of responding to student writing is inherently complex and unavoidably messy. What is more, teachers don't offer their responses in the rather depersonalized manner of copy editors, proofing an ever-flowing stream of comparatively anonymous texts. Instead we experience the work of feedback as a practice *deeply embedded in the contexts of our daily work with the student-authors themselves*.

It is because of this contextualization that we might do well to consider improving teachers' feedback practices not only by sharpening their cognitive awareness of what constitutes good feedback, but also by bringing to bear lessons learned in the psychology of interpersonal relations. In *The Psychology of Interpersonal Relations* (1958), Heider posits a world in which the "common-sense or naïve psychology" implicitly understood by people in their everyday lives might act as the ground for a formalized conceptual framework serving as the "prerequisite for efficient experimentation" (p. 4). By reducing to symbolic terms the complex language people use to describe their experiences with each other, Heider offers to the social sciences community access to a field of "fruitful concepts and hunches for hypotheses [lying] dormant and unformulated in what we know intuitively" (pp. 5-6). Heider's work as a whole deals with several underlying characteristics pertaining to interpersonal relations—e.g., *perception*, *action*,

desire and pleasure, *values*—but it is his seventh chapter, on sentiment, that may be of greatest value to the current study.

Heider's chapter seven focuses on sentiment—"the way that person *p* feels about or evaluates something," that something being either "another person, *o*, or an impersonal entity *x*" (p. 174)—as a driving force in interpersonal relationships. According to Heider's schema, sentiments function either as *likes* (L) or *dislikes* (DL). For example, person *p* may like person *o* (*p* L *o*), or dislike him (*p* DL *o*). Given that nothing in life is simple, Heider also contextualizes these likes and dislikes within a world of other persons, events, and states somehow attached to person *o*, such that when person *p* considers her feelings for person *o*, she might in fact be considering *o* with respect not only to *o* as another self, but also to *o* as a person somehow connected with sentiments of his own, or as a person who somehow identifiably *belongs to* an identifiable group, action, possession, or context. Heider describes this "belonging together" as unit formation (U) (p. 176). Person *o* may for example be a Texan (*o* U *tx*), may be a writer of poetry (*o* U *poetry*), or perhaps be in a longstanding relationship with high college sweetheart (*o* U *cs*). Presumably, *o* would have sentiments about these, as well (e.g., *o* L *tx*, *o* L *poetry*, *o* L *cs*).

At any rate, when *p* considers her feelings about *o*, she very likely takes into account these contextualizing relationships as well, each of which can tend either to confirm or interfere with *p*'s feelings about *o*. If, for example, *o* is a Texan, *p* will consider not only her impressions about *o* but also *o*'s membership among the identifiable group of people from Texas. And in this, the triad of relationships involving *p*, *o*, and *o*'s membership as a Texan can work in any of several ways:

1. *p* likes *o*, and she also likes Texans.
2. *p* likes *o*, but she doesn't really care for Texans at all.

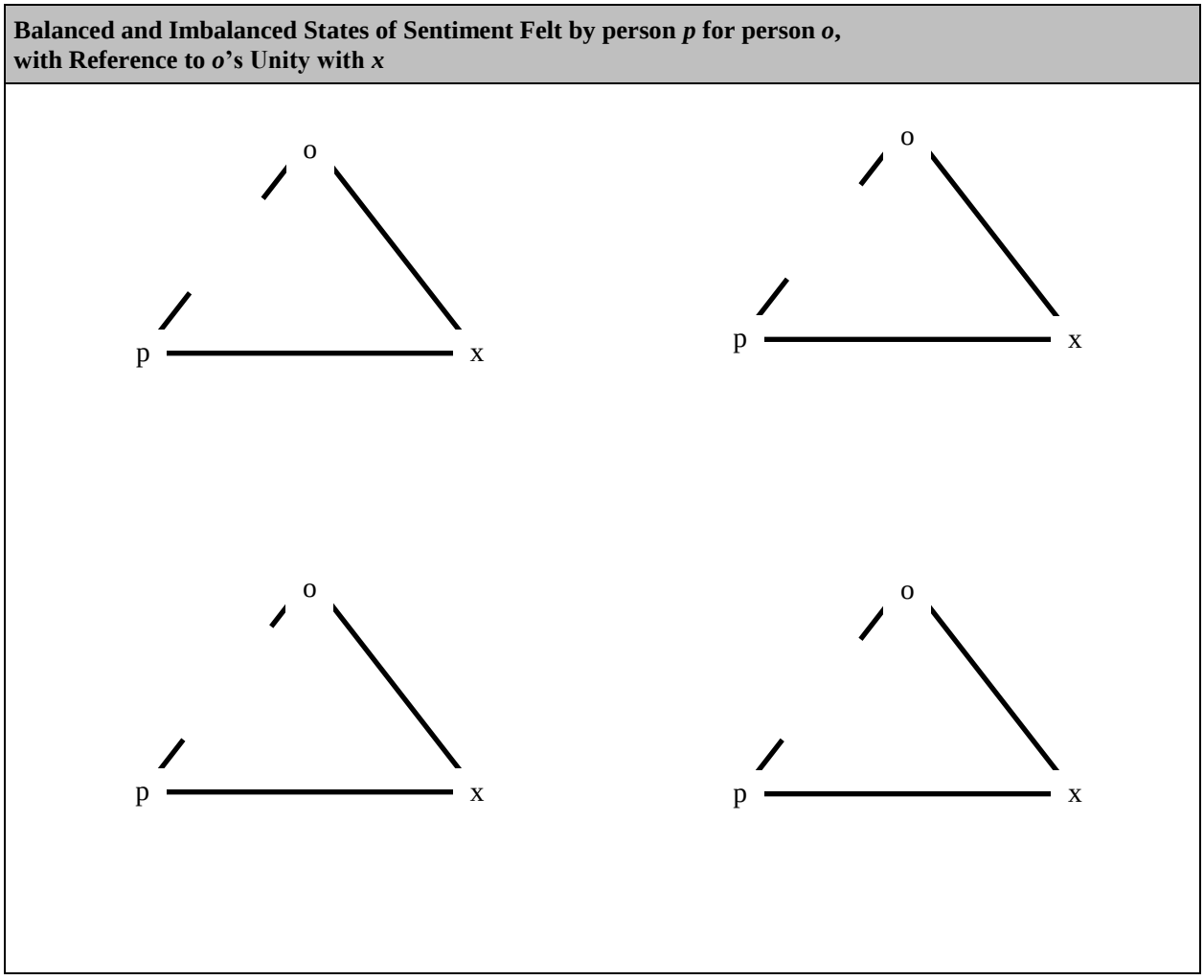
3. *p* dislikes *o*, and she also dislikes Texans.
4. *p* dislikes *o*, but she thinks Texans are great.

Heider renders these sorts of relationships in a graphical format, much as in Figure 9 below, in which he claims that some of these produce a balanced state in *p*'s sentiments regarding *o*, while others produce an imbalanced state. The upshot of such triadic relationships—and a key understanding for the purposes of this study—is Heider's claim "that sentiment relations and unit relations tend toward a balanced state" (p. 201), which is to say that *p* will feel a degree of cognitive and emotional dissonance until she lands on one of two basic options:

1. That her feelings for *o* align (positively or negatively) with her feelings for Texans in general, so that *p* feels more favorably about *o* on account of his identification as a Texan; or that she decides that Texans really can't be all that wonderful, on account of her sentiments about *o*.
2. That, perhaps, she begins to think of *o* in a more complex, fragmentary sort of way such that she holds on to some qualities of *o* as likable (or not) *despite* his continued identification as a Texan, much as in Figure 10. Of course, the problem with this choice is that now there is a lack of balance—and a concomitant "stress toward change" (p. 201)—in how *p* sees *o* himself.

So what does this have to do with teachers, students, and the offering of feedback to student essays? If Heider's interpersonal theory is meaningful, it would be reasonable to argue that a triadic relationship exists between a teacher (*t*), student (*s*), and essay *(e*), such that the teacher either likes or dislikes the student (*t* L *s*, *t* DL *s*); and that the essay is seen as belonging to the student, such that student and essay form a unit (*s* U *e*). Assuming teachers who like their

Figure 9 Triadic Interpersonal Relationships, Adapted from Heider (1958)



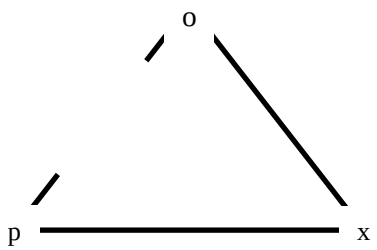
students,[15] we are presented with two possibilities; either the teacher likes the student and also approves of the student's essay, or the teacher likes the student but disapproves of the essay,

[15] While it would be naïve to assume that all *real-world* teachers like all of their students, for the purposes of this illustration, such will nevertheless be the assumption for the sake for this argument. If teachers do not like their students, our work toward achieving first-rate feedback may represent a futile struggle.

leading to an imbalanced state and an internal "stress toward change," that is toward reconciling the good feelings for the student with the disapproval of the student's work.

Figure 10 Resolving a "Stress toward Change," Adapted from Heider (1958)

The Resolution of an Imbalanced State in *p* with Respect to the Conflict She Feels between Her Positive Sentiment toward *p* and Her Negative Sentiment toward *tx*



Certainly, *both* of these scenarios occur for English language arts teachers on a regular basis. While many of our students write well under at least some of the conditions we set for them, most of our students are still in need of tremendous support along the path to writing maturity. Hence the need for their continued education, hence also the need for the feedback that we offer. It is thus probably not a stretch at all to conceptualize the teacher's response to a liked student's poorly written essays much in the way suggested in Figure 10, where *p* likes *o* despite *o*'s identification as a Texan, choosing in fact to dissociate *o* from *tx* to whatever degree is possible. Similarly, we like the student, and we try not to focus too hard on identifying the student to heavily according to the comparative incompetence of his written work.

Yet here's the rub.

English language arts teachers don't simply develop working relationships with our students one at a time. If we did, our work as feedback providers would be easy. On one hand the evaluative task itself would represent a negligible burden on our time or energy, laughably short of the eighteen hours of reading suggested in chapter one. On the other hand, our concentrated interaction with the solitary student would itself be a powerful force inducing a positive relationship. "The tendency," claims Heider, "is for *p* to like a person with whom he has contact through interaction or proximity" (p. 188). With only one student per school day, our interaction with the individual student might complete overwhelm any frustration we feel about his writing. Such—for that matter—is the good parent's case about any of his own child's failings; we *love* the child, and do not fret overmuch when the child reads poorly, fails to win first place in a competition, or even misbehaves. Taken against the great depth of the parent-child relationship, these are trivial matters.

Similar issues—e.g., students' failure to thrive academically or behaviorally in the classroom—are not, however, so trivial for teachers saddled by overburdening caseloads; the situation of having our limited attention distributed across large numbers of students attenuates our ability to interact with individuals frequently enough so as to promote the development of richly rewarding interpersonal relationships that might offset our frustrations with their so-called failings. All too often we are more probably neutrally or even negatively disposed to even our "likable" students, failing to distinguish each one of them clearly as a *David,* or an *Anna,* or a *Trevor*—each with a richly complex life-story—and thus we slip too easily into reductive perspectives of each child as an *s*, the author of an essay, *e*, whose evaluation will be more of a

burden to endure than an opportunity to savor. Our students become numbers, or variables, as it were.

And so what we may need in this recessional era of overcrowded classrooms is not so much another cognitively oriented innovation so much as a psychologically driven one. We need a way to induce a greater sense of interpersonal warmth between teacher and student to counterbalance if not overcome the dread we feel each time we face a stack of essays-to-be-reviewed. And perhaps we can achieve this by returning to Heider's theory to recall what sorts of dynamics are working for us already, to beware those working against us, and to tap into two that can specifically be put into service for the improvement of our feedback practices. Of the forces working toward the strengthening of teacher-student relationships are that teachers frequently see themselves as similar to their students, having themselves endured and succeeded at obtaining an education ("*p* tends to like a similar *o*," p. 184). Teachers also enjoy the benefit of working with students over a long enough time as to develop a sense of familiarity with them ("*p* tends to like a familiar *o*," p. 192). And whenever teachers see their work as educators as oriented toward benefiting their students, they also tend to like their students on the basis of this sense of their work's purpose ("*p* benefits *o* induces *p* likes *o*," p. 199).

Yet despite these forces toward strong interpersonal relationships, we do face a at least a few challenges anticipated by Heider's work. Each year, for example, having guided a cohort of students toward meaningful progress in their education, we receive a new group of "knuckleheads," who aren't nearly as competent as last year's students ("*p* tends to dislike an unfamiliar *o*," p. 193). And within each new class, there are undoubtedly a few "cherubs" with whom it will be hard to work . . . and on account of this, we will tend not to work as well with them as with our star students ("*p* dislikes *o* induces *p* avoids *o*," p. 191). Moreover, we are

prone—particularly as we grow older and more competent in the content areas we teach—to see ourselves as essentially quite different from our students. "Kids these days" we'll sometimes say to each other, "just aren't what they used to be" ("*p* dissimilar to *o* induces *p* dislikes *o*," p. 186).

Yet the above-mentioned competing forces toward better or worse interpersonal relations between ELA teachers and their students all serve merely as background and context with respect to the point toward which this discussion has been heading now for a few pages: Offering feedback to students is not merely a cognitive task whereby we illuminate to students their comparative successes and failures in the skills pertaining to essay writing; it is additionally an interpersonal transaction with the potential to induce in students the psychological desire to improve their outcomes and close the gaps between current levels of success and intended targets. And if this is so, if we thus recall that that feedback has already been considered along lines congruent with its contextualization within the realm of interpersonal psychology, why ought we not consider not only the effects that teachers' comments have on students but also the effects that might arise from students adding "feedback" of their own to papers before their submission? Such comments could be contrived in such a way so as to tap into two of Heider's theoretical planks rather nicely:

First, were teachers to set up a pre-submission routine wherein students were encouraged to examine a series of specifically teacher-prompted (even teacher-modeled), last-minute "quick fixes" pertaining to Standard Written English and formatting conventions, they might have put themselves into the position of feeling not only that they have done their students a good deed ("*p* benefits *o* induces *p* likes *o*," p. 199), but also that the students' light editing itself would constitute a good deed done by the students on behalf of their teachers ("*o* benefits *p* induces *p* likes *o*, or *p* tends to like a person who benefits him" (p. 199). For example, in the world of a

senior-level ELA teacher—where the noting adherence to MLA conventions on senior research papers is a high-value target but also a time-consuming, mind-numbing chore—the gift of time saved by not having to mark every little error would be a significant one, one that would tend to diminish the teacher's dislike of reading essays while increasing her sense of liking the students themselves. Second, were this pre-submission routine also to incorporate questions about the papers directed specifically to the teachers from the student-authors themselves, the routine would tend to increase the proximity of teacher and student by way of the student's written voice reaching out rather directly to the teacher as another individual, not simply to the teacher as a generic evaluator ("*p* in contact with *o* induces *p* likes *o*. The tendency is for *p* to like a person with whom he has contact through interaction or proximity," p. 188).

For these reasons, a central pursuit of this project is to determine whether a set of simulated "student-authored" comments applied to a ready-for submission essay will induce in teachers a greater tendency to provide comments more closely approximating those deemed most effective by the existing research in optimal feedback practices—not because the teachers will have been coached in the best practices of offering feedback, but simply because they will have been responding in kind to the kindness of our students. In other words, in the hypothetical world represented by this study's simulated student essays, the teachers will have tricked themselves into better habits of feedback—even if they haven't had time to stop and consider why.

Chapter Summary

General-impression marking (Cooper, 1977) is a subset of holistic scoring frequently used in the classroom, in high-stakes testing, and under research conditions for the purposes of sorting and grouping generalized outcomes. Sometimes we use general-impression scores to sort student performances (e.g., essays or exams); at other times we are more interested in the what

these scores say about the students' institutions or even about the scorers themselves. It is important to remember, though, that general-impression scores are "totalizing," that they do not give us information about the component characteristics of a performance.

Feedback (Brookhart, 2008; Hattie & Timperley, 2007) is among the strategies that teachers and students can use to reduce the gaps between students' current understandings/performances and their desired goals. Feedback answers three essential questions—*Where am I going? How am I going?* and *Where to next?*—and it does so by various combinations of content characteristics: *focus* (on the task, the process used to complete the task, the students' self-regulation, or the student's individual self), *comparison* (to the criteria for success, to the student's previous attempts, to other students' outcomes), *function* (description or evaluation), *valence* (positive or negative), *clarity* (clear or unclear), *specificity* (nitpicky, just right, overly general), and *tone* (communicating/not communicating respect for the student). Generally speaking, feedback is among the most powerful of school-applied interventions, with an effect size of 0.79, as compared to the 0.40 effect of schooling in general (Hattie, 1999).

Given teachers' lack of time to adequately investigate and reflect upon the research-proven attributes of highly effective feedback, however, it may not serve their students well simply to dictate that teachers memorize and implement a rubric for optimal practice. Instead, we might draw from Heider's (1958) theory of interpersonal relations to explore a minor shift in teachers' evaluation and feedback routines, whereby teachers might delegate part of the evaluative process to students themselves, gaining by such a delegation not only an immediate time-savings but also a subtle investment in the interpersonal relationship between teacher and student, such that teachers will naturally—not by cognition but by desire—tend to provide higher-quality feedback than would otherwise the case.

CHAPTER THREE: METHODS

The participants, instruments, and procedures used in this study appear below, as does a description of the data analyses to be formed to explore the study's proposed research questions.

Participants

Sampling Plan and Characteristics of the Represented Population: The study participants comprised the grade 6-12 English language arts teachers (68 middle school teachers, 60 high school teachers) of a suburban school system in Johnson County, Kansas. Teachers in this district serve a comparatively high SES population, with only about 5% of the student population considered "economically disadvantaged." As might be expected from such an affluent district, students in this district achieve high scores on various mandated assessments. In 2009, 95.8% of the district's 11th-grade students scored at or above state standards on the Kansas Reading Assessment, compared to a state average of 84.3%. Similarly, in the 2009 Kansas Writing Assessment, 88.8% of the district's 11th-grade students scored at or above the state standard, compared to a state average of 71.7%. The teachers themselves also exceed state averages. 96.24% of the district's English language arts teachers (middle and high school) are considered "highly qualified," compared to only 94.34% of English language arts teachers statewide (Kansas State Department of Education, 2009).

From one perspective, these study participants represent a convenience sample, as they are the my colleagues. Yet *because* they are my colleagues, they also represent a sample of particular interest: I am interested to learn if my research idea bears quantitative and qualitative fruit within my own teaching community. In other words, if I can use this study to introduce an advantageous practice into my immediate academic culture—or even merely to spur further thinking in this direction—my work in this project will have succeeded, regardless of my

hypotheses' outcomes.

Prior to the data collection procedure, these teachers were advised of the nature and parameters of the study. Their written consent was obtained. Because the data collection activity was embedded in a district-sponsored professional development session, the teachers received professional development points for their participation in the study. Their participation in the study required one hour of time, including the grading/feedback task itself and the professional development presentation that followed.

Human Subjects Issues: This study's data set includes written samples of teacher work, as well as survey responses and follow-up feedback from these same teachers. Human subjects approval was sought from the Institutional Review Board at the University of Kansas and from the host district's research-ethics gatekeepers.

Group Design: Participants were randomly assigned into each of two subject groups according to the text versions they received (described below in Instrumentation) as well as the order in which those versions appear in their packets. Group 1 received "clean" copies of the texts; group 2, "annotated" copies. Within each group, half the participants' essays were in the order *A* ("well composed"), *B* ("uneven results") with the order reversed for the other half.

Setting

Primary data collection took place on the morning of October 15, 2010, in the "commons area" of one of the host district's high schools. Although this was an open space, the circumstantial context of a professional development day reduced distractions to a reasonable level. The participating teachers arrived and received materials as described in the procedures section below, and seated themselves at six-top tables throughout the room. Teachers arrived with the knowledge that they would be involved in a two-hour professional development session,

the first part of which would be devoted to collecting data for research being conducted by a school district employee who was pursuing an advanced degree in English education.

Instrumentation

Simulated Texts for Evaluation by Teachers: According to Lynch and Klemans (1978), the "ideal vehicle" for communicating to student their papers' strengths and weaknesses is a face-to-face conversation: "As one student put it, the most helpful comments are those 'spoken to myself, and not comments that are written down on a paper.' Time constraints and class size, however, often force the teacher to rely heavily on the written comment." (p. 180). It is for these real-world limitations that the current study's comprised two essay sets simulating the work of two eighth-graders and two twelfth-graders, to which participating teachers were to offer their written comments.

Middle School Texts: Middle school teachers received a pair of papers written in response to an untimed prompt for purpose-oriented personal narrative—a major text type for eighth-grade students. For the written response, students were to recall a "specific lesson they've learned" in or out of school, and to write with details about the lesson, its teacher, and the importance they attribute to the lesson learned. Each text was one double-spaced page in length.

One of the prepared texts (*MSA*) represented a well-composed response with respect to the stated targets for the assignment, written within the hypothetical context of recent lessons in narrative writing, comma usage, and paper formatting, as well as an assumed background in which developmentally appropriate demonstrations of rich content, clear organization, interesting sentence fluency, effective word choice, and reasonable adherence to conventions should be possible. For the purposes of grade-assignment, teachers were advised to consider the domains of *content*, *organization*, *sentence fluency*, *word choice*, and *conventions* in light of the

host district's teachers' habituation with Education Northwest's 6+1 Trait model of evaluation (Education Northwest, 2011), a model which these teachers employ biennially when serving as readers for the state-mandated Kansas Writing Assessment (KSDE, 2008), and which many of them employ in their daily work with students. To avoid a potential ceiling effect, Text *MSA* did not represent a perfectly accomplished sample of writing, but one whose strengths far outweighed its weaknesses.

A second text (*MSB*) was similarly composed, but in such a way as to demonstrate comparatively uneven results across the highlighted domains, achieving reasonable success in some areas of the guidelines while leaving considerable room for improvement in others.

A "clean" copy of each prepared text was preserved. Then an "annotated" copy of each was created to simulate the outcomes of a self-evaluation process capable of being administered to students, as described in the Purposes of the Study section of Chapter One. The annotated copy for each prepared text successfully identified and corrected many but not all of the formatting errors introduced into the prepared texts, while incorrectly introducing another two or three meaningful errors by way of faulty annotations. Additionally, the annotated copy provided two author-written questions directed to the teacher about the paper itself and/or its composition. Thus "clean" and "annotated" versions were prepared for each of the two simulated texts (See Appendices A-D for MSA_{clean}, $MSA_{\text{annotated}}$, MSB_{clean}, $MSB_{\text{annotated}}$).

35 copies were made for each of the four essay versions. Copies of MSA_{clean} and MSB_{clean} were placed together into one set of evaluator packets. In half of these packets MSA_{clean} appeared before MSB_{clean}; in the other half this order was reversed. Similarly, copies of $MSA_{\text{annotated}}$ and $MSB_{annotated}$ were placed together into another set of evaluator packets. As with the first set, in half of these packets $MSA_{\text{annotated}}$ appeared before $MSB_{annotated}$; in the other half this order was

reversed. After additional contents were added, as described below, these packets were sealed and randomized in preparation for distribution to the participating teachers at the beginning of the data collection cycle.

High School Texts: High school teachers also received a pair of papers written in response to an untimed prompt for one paragraph describing either (a) the research skills the author had learned best during the previous five weeks of study or (b) the research skills remaining to be learned deemed by the author as most necessary to be learned before beginning a research paper in the following academic quarter.

One of the prepared texts (*HSA*) represented a well-composed response with respect to the stated targets for the assignment, written within the hypothetical context of recent lessons in (a) the use of online library catalogs and databases for research-appropriate resources, and (b) the use of MLA conventions for documenting print and electronic sources (Modern Language Association, 2009), as well as an assumed background in which developmentally appropriate demonstrations of rich content, clear organization, interesting sentence fluency, effective word choice, and reasonable adherence to conventions should be possible. As with the middle school texts, these domains were selected in light of the host district's teachers' habituation with the Education Northwest 6+1 Trait model of evaluation (Education Northwest, 2011). To avoid a potential ceiling effect, Text *HSA* did not represent a perfectly accomplished sample of writing, but one whose strengths far outweighed its weaknesses.

A second text (*MSB*) was similarly composed, but in such a way as to demonstrate comparatively uneven results across the highlighted domains, achieving reasonable success in some areas of the guidelines while leaving considerable room for improvement in others.

A "clean" copy of each prepared text was preserved. Then an "annotated" copy of each

was created to simulate the outcomes of a self-evaluation process capable of being administered to students, as described in the Purposes of the Study section of Chapter One. This annotated copy for each prepared text successfully identified and corrected many but not all of the formatting errors introduced into the prepared texts, while incorrectly introducing another two or three meaningful errors by way of faulty annotations. Additionally, the annotated copy provided two author-written questions directed to the teacher about the paper itself and/or its composition. Thus "clean" and "annotated" versions were prepared for each of the two simulated texts (See Appendices E-H for HSA_{clean}, $HSA_{annotated}$, HSB_{clean}, $HSB_{annotated}$).

30 copies were made for each of the four essay versions. Copies of HSA_{clean} and HSB_{clean} were placed together into one set of evaluator packets. In half of these packets HSA_{clean} appeared before HSB_{clean}; in the other half this order was reversed. Similarly, copies of $HSA_{annotated}$ and $HSB_{annotated}$ were placed together into another set of evaluator packets. As with the first set, in half of these packets $HSA_{annotated}$ appeared before $HSB_{annotated}$; in the other half this order was reversed. After additional contents were added, as described below, these packets were sealed and randomized in preparation for distribution to the participating teachers at the beginning of the data collection cycle.

Additional Contents of the Teacher/Participant Packets: In addition to the two simulated texts, evaluators received the following documents:

The Assignment Contexts, Prompts, and Performance Targets: Accompanying the texts described above were prompts making clear (a) the instructional context of the simulated papers, and (b) that the papers have been composed in a writing process involving direct instruction, the drafting of brief but well-crafted texts, feedback, and revision. Middle school (Appendix I) and High School (Appendix J) variants were provided.

A "Teacher-Created" Exemplar: This document represents a meaningful example that the hypothetical teacher would have provided to students as an *authentic, teacher-generated* response to the prompt. It is not an attempt by the teacher to "recreate" the feel of a student essay but rather an attempt to show students that the teacher is engaged in learning processes similar to their own and that writing about these learning processes is not only "meaningful" for students but also for adults. Middle school (Appendix K) and High School (Appendix L) variants were provided.

Scoring and Feedback Instructions: The participating teachers were asked to read both written responses, applying a percentage grade to each, as well as providing comments to justify this grade and to help the paper's author improve a subsequent draft of the text. To aid in their assessment and feedback activities, teachers received information regarding the paper's hypothetical context within ongoing lessons specifically focused on narrative writing, comma usage, and paper formatting; they were encouraged, too, to consider matters of content, organization, sentence fluency, word choice, and conventions.

While the control-group scorers of the 8^{th} and 12^{th} grade papers were asked simply to "give the paper the grade it actually deserves" at the current stage of revision, according to the scorer's "honest standards" (Appendix M) experimental-group scorers received the following additional instructions: "Where the author of this paper has already made corrections to the essay or asked meaningful questions, feel free to credit the existing annotations to the student's grade . . . As you make comments, do not feel the need to repeat what the student has noted already" (Appendix N).

Procedures

Distribution of Evaluator Packets, Orientation to the Task: Upon entering the study site,

teachers received the randomly sorted, sealed packets and listened to a brief orientation about the simulated assignment. Once this orientation was complete, teachers opened their packets, read the included consent forms (Appendix P), and signed these if they were willing to participate in the study. Next, their attention was drawn to the various packet contents as described above, with particular emphasis given to observe carefully the "Your Tasks as Evaluator" section. It was explained to teachers that they had already been randomly selected into two groups and that the "Your Tasks" section represented the condition of scoring and feedback that was unique to their group. Time was allowed for the participants to ask questions of the principal investigator and/or the District Coordinating Teacher for Communication Arts, who had assisted in organizing the professional development session and who had volunteered to help administer the scoring procedure.

Once these questions had been resolved, three minutes were allowed for participants to review the context and prompt, the performance targets, the evaluator tasks, and the exemplar essay. At the end of this session, teachers were again reminded to observe the shaded "IMPORTANT" section of the "Your Tasks as Evaluator" section. Then followed two six-minute blocks of time to score and evaluate the two student essays in their possession. At the end of these blocks of time, the participants were asked to complete any comments they were currently writing, to ensure that they had provided a percentage score, and then to discontinue any further evaluative work with the essay.

Debriefing and Professional Development Session: Once the scoring session was complete, participating and non-participating teachers received a debriefing regarding the study's aims. Further, they received an introduction to the self-evaluative process suggested by the study's simulated texts. It was explained that a simulation of this process was being tested to

demonstrate whether it had the tendency to shift teachers' attention away from simple edits and sorting-oriented feedback toward a focus on feedback that involved a richer assortment of the seven feedback characteristics described by Brookhart (2008)—*focus*, *comparison*, *function*, *valence*, *clarity*, *specificity*, and *tone*.

Further, the presentation suggested various benefits that might accrue to our student writers should we adopt the strategy of asking them to provide predicted scores and self-evaluative comments prior to submitting essays for teacher grading and feedback. Schunk (2003) has claimed that positive self-evaluations are "critically important for maintaining self-efficacy for learning and performing well" (p. 164), while Bandura (1986) has demonstrated that even poor self-evaluations are not an obstacle to self-efficacy or motivation as long as students maintain the belief that they can eventually succeed once they have adopted new approaches to their work. And lest we fool ourselves into believe that self-evaluation is a task beyond our middle- and high school students' abilities, Andrade, Du, and Wang (2008) have shown it to be an effective practice even among elementary school writers.

Near the conclusion of this presentation, teachers were asked to review and provide written comments and/or revisions to their own feedback practices as demonstrated on the simulated student essays. These comments and revisions were to be provided in a different color of ink from the original comments, in order to allow the study's coding process not to be muddled by the blending of what function as the two distinct data sets of original versus revised and/or self-reflective annotations.

Data Analysis

Research Question One: A series of independent-samples *t*-tests ($\alpha = 0.05$) determined whether participating teachers in the experimental condition fell prey to so-called grade inflation

in their general-impression scoring with respect to either the stronger (*MSA*, *HSA*) or weaker (*MSB*, *HSB*) papers under their review (H1). To some, such a rise in grades might seem the result of watered-down expectations, in that teachers in the experimental condition were allowed to "credit [a student's] existing annotations" as meaningfully correct in the prepared drafts (*annotated* versions). From the perspective of this study, however, the differences could be taken to indicate that the experimental texts' authorial self-corrections might have interfered with some teachers' tendency to penalize papers disproportionately for errors that are easy to spot and/or that frustrate reading (Chase, 1983; Marshall, 1967).

Research Question Two: Because experimental-condition participants were informed of and perhaps somewhat guided by the simulated student grade predictions and self-evaluative comments—but also because the experimental texts' authorial self-corrections might have interfered with some teachers' tendency to penalize papers disproportionately for errors that are easy to spot and/or that frustrate reading (Chase, 1983; Marshall, 1967)—it was expected that their general-impression percentage scores might more closely agree with each other on the "higher" and "lower" texts than in the control condition; that is to say, their *interrater reliability* would be higher than in the control group (H2). Cherry and Meyer discuss interrater reliability as "the reliability with which raters assign scores to written tests" (p. 33); by way of a concrete example, interrater reliability describes the tendency for writing assessors to assign the same relative values to the various essays in a stack of papers—perhaps not the same scores to the essays individually, but at least the same ranking of papers and the same comparable "distances" separating each from the others. Interrater reliability operates in concert with test consistency and consistency in student performance for a composite construct of instrument reliability. While a holistic concern with instrument reliability is fundamental to meaningful inferences about

student scores, the current study has focused only on the issue of interrater reliability.

The *intraclass correlation* (ICC)—as explained by Shrout and Fleiss (1979) and Cherry and Meyer (1993)—was used to detect levels of interrater reliability of these general-impression scores. Accounting for interrater reliability by way of the ratio of the variance of interest over the sum of the various of interest plus error (Shrout & Fleiss, 1979), ICC has been described by Cherry and Meyer (1993) as "especially appropriate for holistic scoring" (p. 45). Among the salient questions in the use of ICC are "(a) whether individual or composite ratings are the measure of interest, (b) whether all raters rate all texts, and (c) whether ratings are considered relative or absolute" (Cherry &Meyer, p. 46). This study has made use of ICC formulas 3a and 3b on account of the following study parameters: every rater in each condition has rated each of the two possible texts, and the ratings themselves are "relative rather than objective" (p. 49) in that there is no outside criterion our outcome to which the scores refer. Formulas 3a and 3b have been provided in Figure 11.

Figure 11 Intraclass Correlation Formulas 3a and 3b (Cherry & Meyer, 1993, p. 49)

Formula 3a (reliability of a single rating)	
r = (MSp – Mse) / [MSp + (k-1)MSe]	r = reliability coefficient MSp = between persons mean square MSe = error mean square k = number of raters

Formula 3b (reliability of summed or averaged scores)
r = (MSp – Mse) / MSp

Research Question Three: Because "a dramatic increase in the overall richness of feedback" is not the same as "a greater number of feedback comments," probing the data for

answers to question three isn't as simple as illuminating changes in an average general-impression score or increased IRR among the individual scores themselves. But given the seven domains of feedback content outlined by Brookhart (2008)—*focus*, *comparison*, *function*, *valence*, *clarity*, *specificity*, and *tone*—several measurable domains of stasis or change in group means seemed likely under the introduction of student-authored, self-reflective comments to the experimental-set versions of *MSA*, *MSB*, *HSA*, and *HSB*. These changes were hypothesized on the grounds that teachers who find themselves responding not only to a written text but also to *the student's own comments and questions about that text* will find their own sense of the work altered in two ways. First, to the degree that they find the student has already provided a predicted score and corrected at least a few superficial "errors," they will be more likely to relax from self-perceived sorting-oriented obligations and concentrate more or providing learning-oriented feedback. And, second, insofar as they accomplish this redirection of their concentration, they will find themselves more likely to "communicate to a person"—i.e., the author behind the text—than merely to be making "random and disparate criticisms of the formal properties of a text" (Fuller, 1987, p. 308).

In each case, the comparison of means was accomplished by a 2 x2 ANOVA ($\alpha = 0.05$), with *clean* and *annotated* comprising the first factor, and higher- and lower-success papers (*A* and *B*, respectively) and comprising the second.

- $H3_A$: Although FT will remain proportionally constant across conditions, the proportion of comments focused on the student's composing process (FP) and self-regulation (FR) will be greater in the experimental condition, and especially so for the weaker papers, provoked by the student's own handwritten self-evaluative comments having been added to the word-processed essays. That is to

say that because of the student's interjected comments about the paper, the participant-teacher will at times turn somewhat from a focus on the text itself and toward the student-commenter with feedback focused on writing processes and/or the student's self-regulating strategies. In all cases, but especially so with the weaker paper, student-authored comments will be received both as a "benefit" by the teacher and as signal of proximity, thus interfering with teachers' likelihood of dealing with the texts impersonally and increasing their likelihood of dealing with the texts as the products of the students with whom work in interpersonally relevant contexts.

- $H3_B$: Comparisons to the criteria for "good writing" will remain proportionally constant across experimental conditions and relative paper strengths, but comparisons to imagined previous and/or successive drafts will increase under the experimental condition—and more notably so for weaker than stronger papers. Comparisons to the norm of other students' work will be minimal and constant across both groups, as teachers will not have access to enough representative texts to form concrete notions about group norms.
- $H3_C$: The proportional amounts of descriptive and evaluative comments will remain constant across experimental conditions and relative paper strengths, as teachers' responses are likely to be similarly descriptive or evaluative regardless of whether they are responding to the student's text per se or to the student's comments about that text.
- $H3_D$: A higher proportion of comments will possess positive valence in the experimental condition and with higher-quality papers, as teachers in both

situations will adopt a model of communication best described as evaluator-to-person rather than evaluator-to-text. This is to say that as teachers respond to better papers and to papers supplemented with student-provided commentaries under the experimental condition, they will more frequently rise above mere valence-neutral language of editorial symbols and simple edits, and into domains of communication that involve a more interpersonally "positive" and engaging manner of describing the text's strengths and weakness.

- $H3_E$: The proportions of comments judged to be "clear" or "unclear" will remain constant across conditions and degrees of paper strength.
- $H3_F$: The proportions of comments judged to be "specific" or "unspecific" will remain constant across conditions and degrees of paper strength.
- $H3_G$: As with valence—a measure of "positive" communication, even when communicating the necessary improvements to a text—the proportion of comments judged to be helpful in tone (respectful, positioning the student as agent) will be greater in the experimental condition and with stronger papers.

Yet before a single ANOVA could be calculated, the data themselves required coding, and this coding process required a trained assistant, in this case an English language arts teacher with five years of classroom experience preceded by roughly a decade in copy editing. This assistant was trained with respect to Brookhart's (2008) seven domains of feedback content characteristics, which was accomplished by way of a shared reading and discussion of Brookhart's monograph, and supported by way of a coding reference for annotating each participant-teacher's commentary set. The coding reference, with a few brief explanations attached, is provided in Figure 12.

Figure 11 Coding Reference Adopted from Brookhart's (2008) Domains of Feedback Content

Coding Shorthand

Preliminary Issues • Blue and red ink for initial and revised positions. • Ignore marks pertaining specifically to grades (*A* or %). • After we reach agreement, we'll record totals.		
Focus	**W**	Work Itself
	O	Process Used to Complete the Work
	R	Student's Self-Regulation
	P	Person
Comparison	**C**	Criterion-Referenced
	S	Self-Referenced
	N	Norm-Referenced
Function	**D**	Description
	E	Evaluation
Valence *Ignore simple edits.*	**↑**	Positive
	↓	Negative
Clarity	**✓**	Clear
	?	Unclear
Specificity	**A**	Appropriately Specific
	~~A~~	Not Appropriately Specific
Tone *Ignore simple edits.*	**H**	Helpful
	~~H~~	Unhelpful

Sample Comments and Applied Codes

- A summative comment at the end of *MSA*$_{annotated}$: **Conclusion is present, but abrupt.** (WCD↓✓~~AH~~): *This comment's FOCUS is on the work itself, communicating an implied CRITERION that conclusions be present and reasonably developed; we judged the comment DESCRIPTIVE, but considered labeling "abrupt" as the basis for a evaluative code; the comment's CLARITY is such that an 8th-grade student should understand what is meant, but there is insufficient SPECIFICITY to illuminate where, exactly, the abruptness lies; because this illumination is absent, we felt that the valence was NEGATIVE; as was often the where poor specificity and negative valence appeared, we judged this comment's tone UNHELPFUL, showing too little overt respect for the student, too little interest in "inspiring thought, curiosity, or wondering" (Brookhart, 2008, p. 34).*

- A simple, no-valence, no-tone edit by way of a conventionally agreed upon diagonal slash to transform an upper-case letter to lower case, near the end of *MSB*$_{annotated}$: **After a few balls, my ~~D~~ad suddenly got it.** (WCD✓A)

The assistant's background as an editor was in fact a great boon to the coding enterprise—with respect both to the coding process itself and also to the various moments in which were able to pause and reflect upon the manifold characteristics of the data we were reviewing. And it was in these moments of professional reflection that both he and I obtained the greatest insights from this research project—insights that I will try to explain in chapter five's discussion, but insights which nevertheless are only to be achieved when one is awash in the

concrete data themselves. As he said one morning when we were roughly two-thirds through the data set, pouring over this data has been one of the more richly rewarding professional development opportunities of our careers.

Once the training sequence was complete, the assistant and I proceeded through the following steps with each participant-teacher's commentary set. First, we separated the comments provided during the scoring/feedback session from those added subsequently during the professional development session; these judgments were usually easy, as most teachers followed the instruction to use a different writing instrument for the subsequent comments. Percentage scores were similarly excluded from consideration in the feedback apparatus. Next, we reached agreements about the boundaries of individual comments—where they began and ended, whether a series of phrases or clauses was to be scored as a unit or to be broken down into discrete utterances. After making these judgments, we numbered the comments to facilitate comparisons of our initial coding results. Having identified the dataset, each of us worked independently to arrive at codes for an entire document before conferring about our interpretations and rationales agreement about revisions before moving on. The actual coding decisions were based on the following rubric:

- *Focus*: Does the individual comment address the work itself (W), the process the student may have used to complete the work (O), the student's self-regulation (R), the student personally (P)? *Multiple foci were possible*.
- *Comparison*: Does the comment compare the work to a criterion (C), to imagined previous or subsequent works by the current student (S), or to the work of other students (N)? *Multiple comparisons were possible*.
- *Function*: Does the comment describe (D) or evaluate (E) the student's work? *If*

any part of the comment was evaluative, the whole comment was to be coded "E."

- *Valence*: Is the comment framed in a positive manner (🡅) so as to "describe what is well done" and "accompany negative descriptions of the work with positive suggestions for improvement" (Brookhart, 2008, p. 6), or is it framed negatively (🡇), pointing out errors without offering guidance? *If the comment represented a simple edit (e.g., to remedy an error in spelling or punctuation), the comment was to be coded as valence-neutral—receiving neither an up or down arrow.*
- *Clarity*: Is the comment likely to be meaningfully understood (✓) to the purported author of the current text, or is it likely to be unclear (?) requiring follow-up explanations from the feedback-provider?
- *Specificity*: Is the comment appropriately specific (A) so that the writer will understand not only the general concept of this feedback-provider's comment, but also how the comment applies to a specific word, phrase, sentence, paragraph, or section of the text; or is the comment not appropriately specific (~~A~~) being either "nitpicky" (p. 6) or overly general?[16]

[16] In the practice of coding, "nitpicky" proved to be a judgment we never made directly. While some feedback providers in our data set clearly operated upon the principle of offering little other than what we described to ourselves as mechanically operating "simple edits" (e.g., inserting commas, correcting misspelled words, etc.), the "nitpicky" nature of such feedback actually showed up in the proportion of simple edits—which we judged to be *valence- and tone-neutral*—to comments for which valence and tone came into play. For example, one scorer of $MSB_{annotated}$ provided eight comments, *all of which* were valence and tone neutral: 1 for overall page-formatting, 1 for faulty capitalization, 1 for the use of a second-person pronoun, 3 for comma errors, and 2 for supposedly faulty sentence constructions (both of which would usually be judged acceptable in most professional-writing contexts).

- *Tone*: Does the comment "communicate respect for the student and the work," positioning "the student as the agent" in a helpful way (H) likely to "cause students to think and wonder" (p. 7); or does the comment show an unhelpful (~~H~~) disrespect for the student, diminishing agency and thoughtfulness? *As with valence, if the comment represented a simple edit (e.g., to remedy an error in spelling or punctuation), the comment was to be coded as tone-neutral—receiving no code.*

Having used this rubric to reach our independent interpretations for each commentary set, we then discussed our findings. Despite somewhat frequent initial differences (i.e., with 577 of the dataset's total 2283 comments, or 25.3%), we almost always achieved agreement before recording our work and moving on (i.e, for 2278 of 2283 comments, or 99.8%). With respect to the five items for which we were unable to reach total agreement, we recorded in our dataset the elements on which we agreed, but omitted those that remained in dispute. It is worth noting that

---

This was, clearly, a nitpicky comment-provider seemingly attuned only to superficial textual elements, one who would likely benefit from additional professional development in responding to student texts.

By contrast, the middle school teacher we deemed to have provided the richest, most meaningful feedback in our dataset provided a total of six comments on MSB_{clean}, only one of which involved the simple correction in capitalization (e.g., revising "my Mom" to "my mom"). The remainder of her feedback provided a highly engaging mixture of comments focused on the work, on the students process, and even on the student's sense of self-regulation. For example, her final comment on read, "I'm really liking this part!! Why? Because you are beginning to demonstrate how being *responsible* made you feel *great*. Wow—super detail. However, we need to know *more* about how your *mom* taught you this." This comment, scored WCE↑✓AH, was typically engaging, and like the rest of the teacher's feedback was clear and specific, possessing a strongly positive valence and helpful tone.

the work of coding proved an iterative process, frequently requiring our revisitation of earlier coding decisions to help us adjudicate new ones, but at other times to revise earlier judgments is light of later ones. A summary of various impressions we formed and ad hoc decisions we made appears in Figure 13.

Figure 13 Questions Requiring Ad Hoc Decisions during the Coding Process

Questions Requiring Ad Hoc Decisions during the Coding Process

Conventions and Irregularities in Coding Teachers' Feedback

- We agreed to code "*simple edits*" (e.g., CS or FRAG) as possessing neither valence nor tone, but merely communicating in a basic, functional way about "correctness."
- Given the high number of teachers who do not allow second person (i.e., "you" statements) in student texts, any circled "you" was coded "WCD✓A" even without a clarifying comment; similar allowances were made for clear misspellings, incorrectly used homonyms, and faulty coordinating conjunctions (e.g., *nor* instead of *or*).
- Because of expectations within the host district and the simulated assignments themselves—with the assumption of model texts throughout the district—non-annotated circles around departures from MLA formatting, or circles merely noting "MLA" were scored "WCD✓A."
- All other circles without explanatory comments were coded WCD?~~A~~, as was the notation "AWK" (i.e., awkward), when the specific causes for awkwardness were not illuminated.
- When provided as part of a rubric-assessed entry (e.g., *Conventions–4/5: Reasonable control of comma rules*), even "descriptive" comments were coded as evaluative.
- In a few cases, comments left unfinished to the degree of being unintelligible were treated as stray marks.
- *Correctness* of the feedback was ignored; more often than we would have wished, we noted faulty feedback demonstrating a teacher's clear lack of expertise about what counts as "good," or even simply "correct" writing.

Interesting Impressions

- Although not attempting to keep track of trends involving the degree of overlap between *valence* and *tone*—and returning to do so after the fact would be a laborious side-track to the questions this study is pursuing—we perceived so a high degree of correlation in the scoring of valence and tone that we weren't sure about these as meaningfully distinct constructs. A rare *negative-valence* (🡻) but *helpful* (H) comment appears on $MSA_{annotated}23$, where the feedback-provider has responded to the student question "Q1: Do I get of track here, or is it good?" with "*off track*." As the comment is responding directly to the student with uptake of the student's own language, we felt that it was (perhaps marginally) demonstrating respectfulness and a desire to support the student's sense of agency; nevertheless, the comment also seemed to merit a code for negative valence, as it points out the error but does nothing to suggest a direction for improvement.
- At times, we were surprised to find ourselves not only being subject to a halo effect when moving through the comments on a particular teacher's work, but actually *depending on it* to help us make inferential decisions about a particular comment within the context of others on the same page.
- We were surprised by the degree to which *positive* evaluative comments seemed to enhance the sense of a feedback-provider speaking to the author rather than to the text (see Fuller, 1987), even while we agree with Brookhart (2008) that *negative* evaluative comments are caustic.
- Even before running the data through SPSS, we were surprised by the weakness of the middle school commentaries in comparison to those provided to the high school papers: There seemed to a comparatively high proportion of "simple edits" in the middle school set, even under the experimental condition—where many feedback-providers seemed to ignore altogether the student-added comments. Additionally the middle school

dataset gave both coders the impression of containing a surprisingly higher number of *errors* in the comments provided.
- That said, among high school and middles school participants alike, some of the dataset's feedback providers were quite masterful at their work, even under contrived conditions such as those employed in this study. Meanwhile, other feedback providers seemed to work hard at the task while nevertheless providing what research has consistently shown to be poor-quality comments.

Irregularities and Conventions for the General-Impression Scoring
- Papers for which a *letter grade* has been applied rather than a percentage grade received the following emendations: A+: 99, A: 95, A-: 91, B+: 85, B:85, etc.
- Papers receiving a ranged grade (e.g., 78-80) were emended toward the given range's median value, with rounding up to the next integer at .5.

Research Question Four: The raw, qualitative data provided by teachers' initial feedback task, as well as their subsequent self-commentaries and/or revisions were reviewed for thematically coherent elements related to the potential effects of the professional development sequence within which the study's data collection piece occurred. These themes will be addressed in the study's discussion section.

Chapter Summary

In order to investigate the questions posed in chapter one, this study is making use of a mixed-methods research model, whereby participating teachers have been randomly placed into control or experimental conditions and then given a grading/feedback task to complete. Following a qualitative coding procedure to capture the characteristics embedded in the teacher comments, these have been recorded into Excel and SPSS spreadsheets for various quantitative analyses to determine whether, in fact, a change in working conditions contributed to (H_1) an inflation in general-impression scores, (H_2) greater interrater reliability, as measured by intraclass correlations, and/or (H_3) meaningful differences with respect to the feedback characteristics, as measured by a series of two-way ANOVA calculations.

Following the completion of this task, participating teachers were also invited to consider

their work in light of best practices in feedback as described by Brookhart (2008), making self-evaluative comments and/or changes to their comments in light of Brookhart's suggestions. The subsequent data provided in these evaluative comments have been counted and sorted into basic descriptive categories, and will be presented in the chapter five discussion of findings.

CHAPTER FOUR: RESULTS

Chapter four presents results in the following sections. First, descriptive characteristics summarize the general characteristics of the participant sample. Second, statistical analyses demonstrate outcomes for the three research hypotheses. Measures of central tendency and dispersion for the participating teachers' general impression scores, intraclass correlations, and 2x2x2 factorial ANOVAs were performed to test these hypotheses.

Descriptive Data for Participant Groups

59 of the host district's 68 middle school English language arts teachers (30 control and 29 experimental) gave permission for their results to be in the study, achieving an 86.7% participation rate. Participation was somewhat lower among high school teachers, with 50 of 60 (25 control, 25 experimental) giving permission, achieving an 83.3% participation rate.

Research Question 1: General-Impression Scoring—Central Tendency, Dispersion

Hypothesis 1: The simulated self-feedback routine will precipitate a mild inflation of the general-impression scores assigned to student work by experimental-group participants

Table 1 provides the study's central tendency and dispersion results for teachers' general-impression scoring across cohorts (MS, HS), papers (higher, lower), and scoring conditions (control, experimental). Of initial interest in these descriptive data is the demonstration that teachers in both the middle- and high school groups were somewhat more demanding of the papers than anticipated in the study design. In each case the "well-composed response"(*MSA*, *HSA*) was expected to receive a middle-*A* from the control group, while the "comparatively uneven results" of the lower papers (*MSB*, *HSB*) was expected to receive a low-*B*. Similarly surprising were the middle school teachers' perceptions of *MSA* and *MSB* relative to each other. While *MSB* was read by participants nearly as planned—as a high *C* rather than low *B* paper,

with scores of 79.30 and 79.10 by the control and experimental groups, respectively—MSA was perceived by participants as a much more problematic text than intended, receiving scores at least one full letter grade below the expected middle *A* (82.30 and 83.90 for the control and experimental groups, respectively).

Table 1 General-Impression Scores: Descriptive Data by Cohort, Essay, and Scoring Condition

Middle School					
Paper, Scoring Condition	Number	Mean	SD	Min.	Max.
MSA_{clean}	30	82.30	6.798	68	94
$MSA_{annotated}$	29	83.90	6.997	70	93
MSB_{clean}	30	79.30	6.276	68	88
$MSB_{annotated}$	29	79.10	6.689	70	94
High School					
Paper, Scoring Condition	Number	Mean	SD	Min.	Max.
HSA_{clean}	25	86.84	6.276	70	98
$HSA_{annotated}$	25	86.16	5.520	74	98
HSB_{clean}	25	71.88	9.139	50	88
$HSB_{annotated}$	25	78.00	6.110	65	93

Table 2 provides the results of four independent samples *t*-tests of these descriptive data, conducted to check for significant differences in the general-impression scores between control and experimental conditions. In the first three of these *t*-tests, no such differences appeared. In the fourth grouping, however, the *HSB* experimental group's 78.0 average did prove significantly higher than the control group's 71.88.

Table 2 General-Impression Scores: Independent Samples *t*-Test Results, One-Tailed

Comparison	t	df	p
MSA_{clean}, $MSA_{annotated}$	-0.889	57	0.189
MSB_{clean}, $MSB_{annotated}$	0.116	57	0.454
HSA_{clean}, $HSA_{annotated}$	0.407	48	0.343
HSB_{clean}, $HSB_{annotated}$	-2.783	48	0.004

Research Question 2: Intraclass Correlations

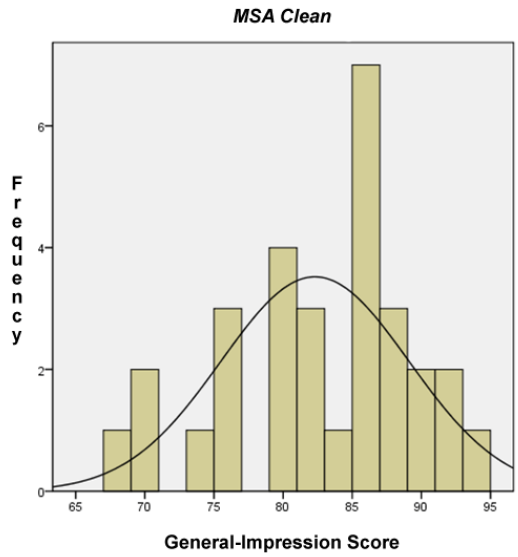
Comparing control- and experimental-group averages allows only a small sliver of light to shine through on teachers' general-impression scoring data. A somewhat richer picture is afforded when we also consider the degree to which teachers agreed upon the papers' relative merits. Figures 14 and 15 introduce this picture graphically, making visible a few scoring characteristics worthy of note.

What is most readily visible are the generally different levels of agreement about scores between the middle and high school levels. On one hand, middle school teachers *as a group* seemed comparatively uncertain about the general-impression scores they gave. This lack of certainty can be seen in the middle school histograms, all of which are somewhat platykurtic or "flat,"[17] with kurtosis values as follows: MSA_{clean} = -0.417, $MSA_{annotated}$ = -1.003,

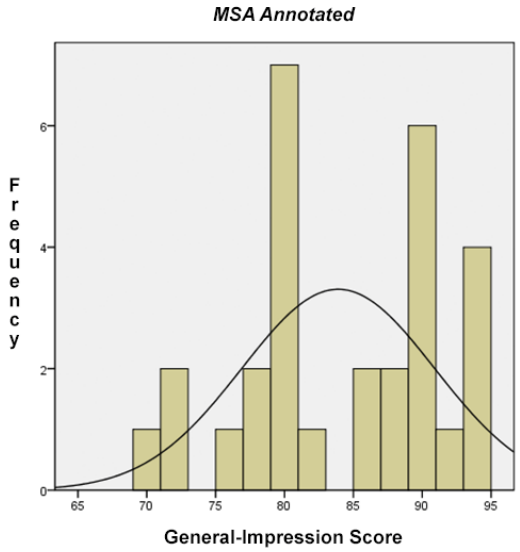
[17] A platykurtic histogram whose values described a *perfectly* flat curve—i.e., a horizontal line—would depict a grading scenario in which every possible score had given equally often to the essay under review, implying no agreement whatsoever among scorers about its relative value. By contrast, to the degree that scorers shared a closely aligned perception about a paper's merits, a leptokurtic curve would take shape, surging upward steeply around the mean score.

Figure 14 Histograms for Middle School Essay Scoring, By Paper and Group

Histograms for MSA_{clean} and $MSA_{annotated}$

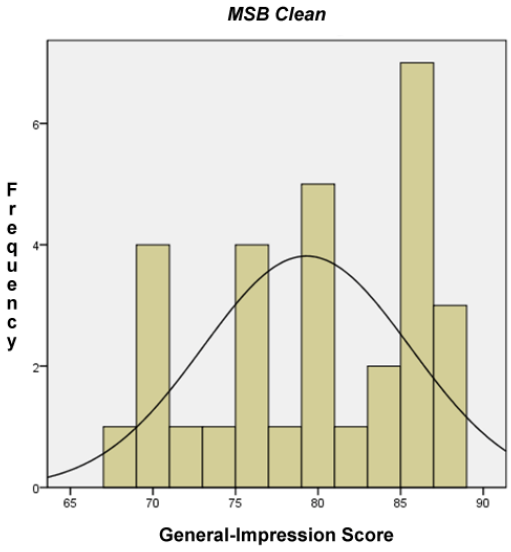


Mean = 82.30, SD = 6.798, N = 30

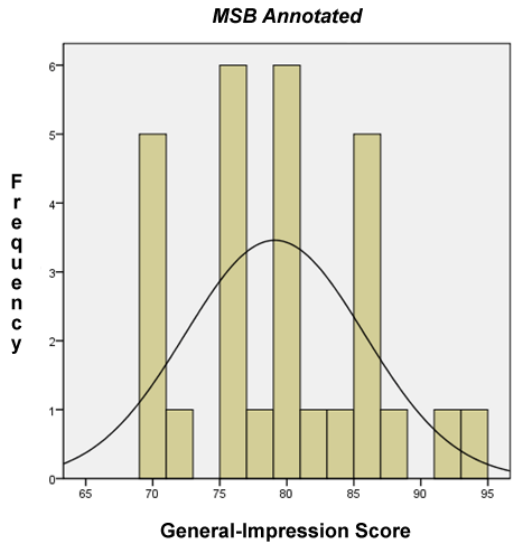


Mean = 83.90, SD = 6.997, N = 29

Histograms for MSB_{clean} and $MSB_{annotated}$



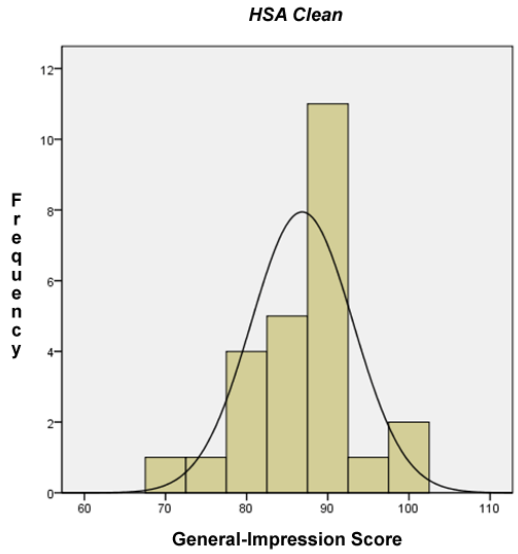
Mean = 79.30, SD = 6.276, N = 30



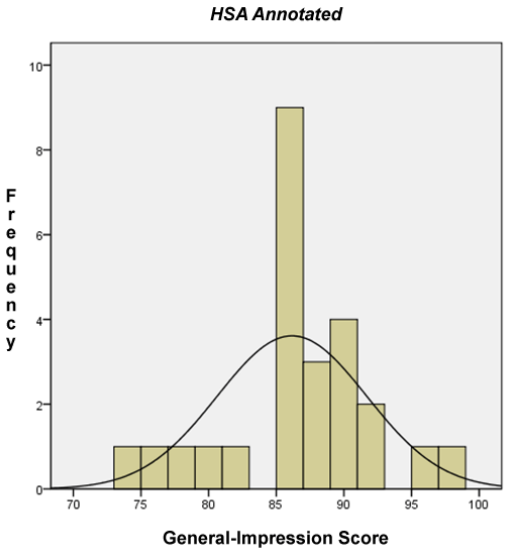
Mean = 79.10, SD = 6.689, N = 29

Figure 15 Histograms for High School Essay Scoring, By Paper and Group

Histograms for HSA_{clean} and $HSA_{annotated}$

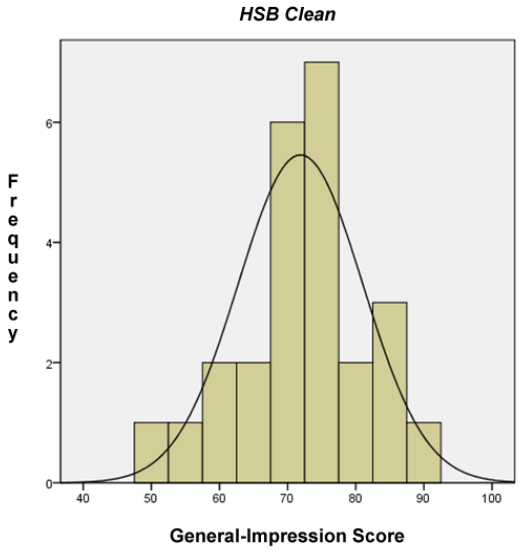


Mean = 86.84, SD = 6.276, N = 25

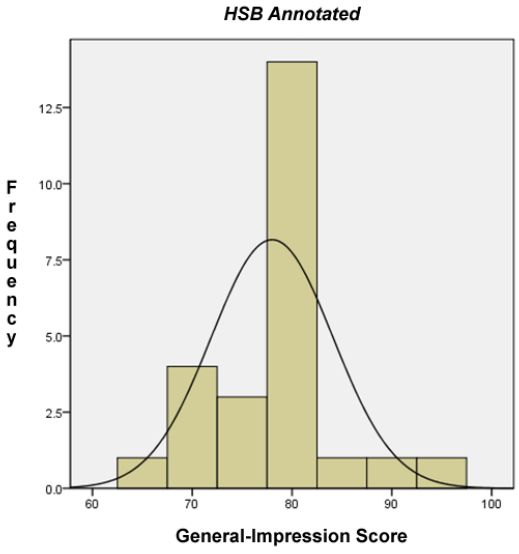


Mean = 86.16, SD = 5.520, N = 25

Histograms for HSB_{clean} and $HSB_{annotated}$



Mean = 71.88, SD = 9.139, N = 25



Mean = 78.00, SD = 6.110, N = 25

MSB_{clean} = -1.223, $MSB_{annotated}$ = -0.610. High school teachers, on the other hand, enjoyed a greater tendency to award similar scores to their two essays, providing for repeatedly leptokurtic values across the scoring conditions, as follows: HSA_{clean} = 1.345, $HSA_{annotated}$ = 0.466, HSB_{clean} = 0.336, $HSB_{annotated}$ = 0.906.

The visual impression created by these histograms is further reinforced by the interrater reliability calculations at the heart of the second research question: *Hypothesis 2: The simulating self-feedback routine will effect a marked improvement in the interrater reliability (IRR) of the general-impression scores assigned to student work by experimental-group participants, as measured by intraclass correlation formulas 3a and 3b.* Table 3 provides the results for interrater reliability.

Table 3 IRR: ICC Formulas 3a and 3b, Variants for Consistency and Absolute Agreement

Paper, Scoring Condition		Consistency		Absolute Agreement	
	Cronbach's Alpha	Single Rating Reliability (3a)	Group Reliability (3b)	Single Rating Reliability	Group Reliability
MS_{clean}	0.672	0.064	0.672	0.066	0.679
$MS_{annotated}$	0.881	0.203	0.881	0.178	0.862
HS_{clean}	0.982	0.681	0.982	0.641	0.978
$HS_{annotated}$	0.984	0.706	0.984	0.491	0.960

As these calculations show, interrater reliability was much higher among high school teachers than middle school teachers across both *clean* and *annotated* conditions. To illustrate, when considered from the framework of consistency rather than absolute agreement—that is, from the tendency of teachers' scores to follow parallel lines, regardless of whether these lines represent the work of "harder" or "easier" graders—the *group* reliability was so low among

middle school control-group teachers (r = 0.672) that it was beaten by the *individual* reliability numbers for both high school conditions ($r_{HSclean}$ = 0.681, $r_{HSannotated}$ = 0.706)

That having been said, hypothesis 2 achieved only mixed results. Among the middle school experimental-group participants, individual and group reliability results were consistently higher than for their control-group counterparts. This improvement appeared both in the reliability results for individual teachers and for the groups as a whole. For example, when considering the *consistency* of scores, experimental-group reliability numbers were 0.139 higher for individual scores, and 0.209 for group reliability. By the harder measure of *absolute agreement*, the improvement was not as large but still appreciable, 0.112 higher for individual scores, 0.183 for group reliability. Considered against the backdrop of a longstanding sense among measurement specialists that interrater reliability scores of 0.80 are sufficiently high for even large-scale, high-stakes work (Diederich, 1974; Hillocks, 1986), the control- to experimental-group improvements in absolute agreement from 0.679 to 0.862 do indeed seem noteworthy.

On the other hand, interrater reliability results for experimental-group high school teachers' scores either remained stable or deteriorated with respect to scores in the control group. Where the control group's scorers agreed strongly enough to achieve single-rating consistency results of r = 0.681 and absolute agreement results of r = 0.641, their experimental-group counterparts only improved marginally in consistency, with a reliability score of r = 0.706 while actually losing ground with respect to absolute agreement, falling to r = 0.491. Similarly, their whole-group reliability numbers increased only from 0.982 to 0.984 in consistency, while falling from 0.978 to 0.960 in the measure of absolute agreement. Nevertheless, with whole-group reliability scores well above 0.950 in both conditions, these losses seem trivial.

Research Question 3: Feedback Characteristics

Initial Considerations: The task of examining this study's feedback dataset encountered an unanticipated problem in that several teachers in the participant pool granted permission for the use of their data but did not seem to participate meaningfully in the feedback-providing task. For example, two evaluated papers submitted by permission-granting participants contained no feedback whatsoever, merely a percentage grade. Clearly, such work would be so unacceptably bare in the real world of a learning-oriented classroom as to be beyond the pale of even minimally conscionable practice. These two papers—and in fact a few other similarly information-poor examples—seemed worthy of removal from the analysis of feedback characteristics. Yet before making any decisions about excluding samples from the feedback analysis, it seemed prudent to consider descriptive statistics for the participant data as a whole. Tables 4 and 5 provide these descriptive statistics, accompanied by a visual renderings of the same in Figure 16.

As can be seen in these numerical and visual representations, the average number of comments delivered during the six-minute scoring periods (10.47) is really not all that great, and in fact samples with as few as 5 comments were within the first standard deviation from this average. The data do, however, seem to make clear a natural "break" suggestive of a low cut-off point for an exclusion/inclusion decision. That break appears between the 7 participants (3.2%) who offered only 0-2 comments and the 17 (7.8%) who provided 3 comments. Figure 14 makes this breaking point visible, with the 2-to-3 comment gap being the largest in the dataset.

In this manner it was determined that the minimum total number of comments required for inclusion in the feedback characteristics analysis would be *three*. At this level, all 25 samples from the high school control group (HSA_{clean}, HSB_{clean}) were included in the study, but only 23

and 24 samples from $HSA_{annotated}$ and $HSB_{annotated}$, respectively. Similarly, all 30 samples from MSA_{clean} and MSB_{clean} were included, but only 26 and 28 samples from $MSA_{annotated}$ and $MSB_{annotated}$, respectively.

Table 4 Descriptive Statistics for Total Feedback Comments per Paper across All Conditions

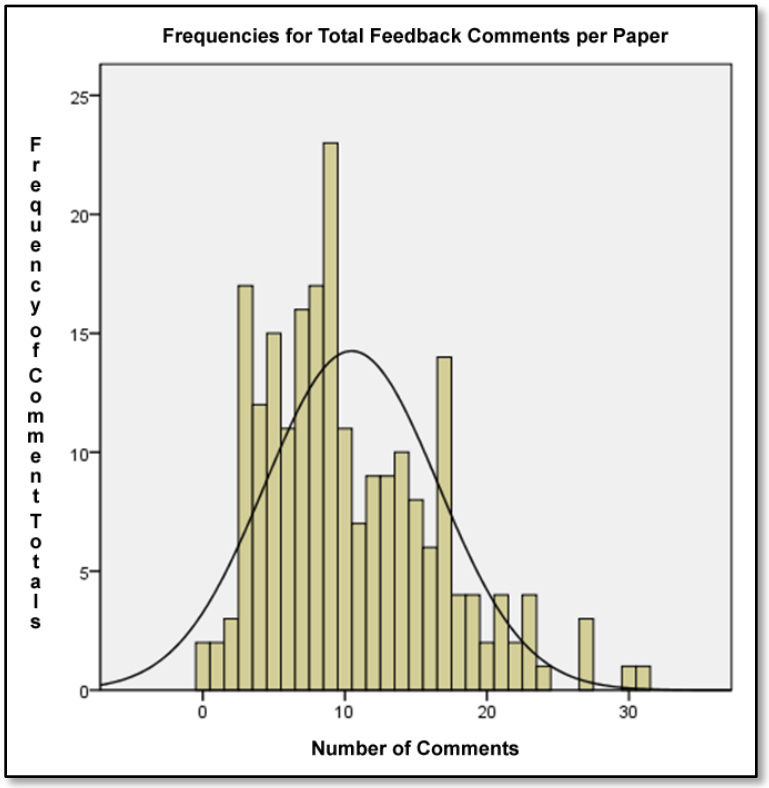
Descriptive Statistics	
Number of Cases	218
Mean	10.47
Median	9.00
Mode	9.00
SD	6.10
Minimum	0
Maximum	31

Table 5 Frequencies for Total Feedback Comments per Paper across All Conditions

Number of Comments	Frequency	Percent	Cumulative Percent
0	2	.9	.9
1	2	.9	1.8
2	3	1.4	3.2
3	17	7.8	11.0
4	12	5.5	16.5
5	15	6.9	23.4
6	11	5.0	28.4
7	16	7.3	35.8
8	17	7.8	43.6
9	23	10.6	54.1
10	11	5.0	59.2
11	7	3.2	62.4
12	9	4.1	66.5

13	9	4.1	70.6
14	10	4.6	75.2
15	8	3.7	78.9
16	6	2.8	81.7
17	14	6.4	88.1
18	4	1.8	89.9
19	4	1.8	91.7
20	2	.9	92.7
21	4	1.8	94.5
22	2	.9	95.4
23	4	1.8	97.2
24	1	.5	97.7
27	3	1.4	99.1
30	1	.5	99.5
31	1	.5	100
Total	218	100.0	

Figure 16 Frequencies for Total Feedback Comments per Paper



$H3_A$: Although focus on the task (FT) and focus on the student's self (FS) will remain proportionally constant across experimental conditions and relative paper strengths, the proportion of comments focused on the student's composing process (FP) and self-regulation (FR) will increase under the experimental condition—and more notably so on the weaker papers—provoked by the student's own handwritten self-evaluative comments having been added to the word-processed essays.

Table 6 provides descriptive sample sizes, means, and standard deviations for each paper in each experimental assignment and feedback cohort, with respect to comments' proportional

focus on the *task* or *text* (FT), on the *process* of composing the text (FP), on the student's *self-regulation* (FR), and on the *student personally* (FS). Table 7 provides summary information for the 2x2 ANOVA on the main and interaction effects for these focal characteristics of teachers' feedback.

Table 6 Descriptive Statistics for Feedback Focus by Experimental Condition and Paper Strength

	Condition	***Weaker***		***Stronger***	
		n	Proportional Mean, SD	n	Proportional Mean, SD
Task (FT)	Control	55	0.969 0.107	55	0.976 0.102
	Experimental	52	0.958 0.110	49	0.948 0.167
Process (FP)	Control	55	0.026 0.064	55	0.012 0.043
	Experimental	52	0.034 0.098	49	0.018 0.055
Self-Reg. (FR)	Control	55	0.032 0.065	55	0.031 0.077
	Experimental	52	0.080 0.148	49	0.115 0.147
Student (FS)	Control	55	0.018 0.135	55	0.006 0.045
	Experimental	52	0.002 0.017	49	0.020 0.143

Table 7 ANOVA Results for Feedback Focus by Experimental Condition and Paper Strength

Source	**Sum of Squares**	***df***	**Mean Square**	***F***	***p***	**Partial η^2**	**Obsrvd. Power**
Focus on the Task (FT)							
Omnibus Model	*0.025*	*3*	*0.008*	*0.544*	*0.653*	*0.008*	*0.161*
Condition (Control vs. Experi.)	0.021	1	0.021	1.379	0.242	0.007	0.215
Paper (Lower vs. Higher)	< 0.001	1	< 0.001	0.012	0.911	< 0.001	0.051
Condition x Paper	0.004	1	0.004	0.267	0.606	0.001	0.081

Source	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial η^2	Obsrvd. Power
Focus on the Process (FP)							
Omnibus Model	*0.015*	*3*	*0.008*	*1.077*	*0.360*	*0.015*	*0.289*
Condition (Control vs. Experi.)	0.003	1	0.003	0.661	0.417	0.003	0.128
Paper (Lower vs. Higher)	0.012	1	0.012	2.525	0.114	0.012	0.353
Condition x Paper	< 0.001	1	< 0.001	0.013	0.910	< 0.001	0.051
Focus on Self-Regulation (FR)							
Omnibus Model	*0.260*	*3*	*0.087*	*6.632*	*< 0.001*	*0.088*	*0.972*
Condition (Control vs. Experi.)	0.233	1	0.233	17.825	< 0.000	0.079	0.988
Paper (Lower vs. Higher)	0.015	1	0.015	1.134	0.288	0.005	0.185
Condition x Paper	0.017	1	0.017	1.304	0.255	0.006	0.206
Focus on the Student (FS)							
Omnibus Model	*0.012*	*3*	*0.004*	*0.406*	*0.749*	*0.006*	*0.130*
Condition (Control vs. Experi.)	< 0.001	1	< 0.001	0.003	0.959	< 0.001	0.050
Paper (Lower vs. Higher)	< 0.001	1	< 0.001	0.045	0.832	< 0.001	0.055
Condition x Paper	0.012	1	0.012	1.185	0.278	0.006	0.192

A medium-sized main effect for the experimental condition was found with FR, $F(1, 207) = 1$, $p < 0.001$, partial $\eta^2 = 0.079$. This result suggests support for certain aspects of $H3_A$. As expected, FT and FS remained constant across experimental conditions and paper strengths. Moreover, the rate of FR increased under experimental conditions from 0.055 (SD 0.115) to 0.070 (SD 0.122). However FR did not increase across paper strength as predicted—neither as a main effect nor under an interaction with the experimental condition. Nor was FP affected in any significant way by experimental condition or paper strength.

$H3_B$: Comparisons to the criteria for "good writing" will remain proportionally constant across conditions and relative paper strengths, but comparisons to imagined previous and/or successive drafts will increase for papers under the experimental condition—and more notably so for weaker than stronger papers. Comparisons to the norm of other students' work will be minimal and constant across conditions and paper strengths, as teachers will not have access to

enough representative texts to form concrete notions about group norms.

Table 8 provides descriptive sample sizes, means, and standard deviations for experimental assignment, paper strength, and feedback cohort, with respect to the comments' proportional engagement with a *stipulated criterion* (C), *the same student's other work* (S), or the *norm of other students' work* (N). Table 9 provides summary information for the 2x2 ANOVA on the main and interaction effects for these elements of comparison within teachers' feedback.

Table 8 Descriptive Statistics for Comparisons by Experimental Condition and Paper Strength

	Condition	***Weaker***		***Stronger***	
		n	Proportional Mean, SD	n	Proportional Mean, SD
Criteria (C)	Control	55	0.971 0.105	55	0.971 0.106
	Experimental	52	0.968 0.107	49	0.955 0.158
Self (S)	Control	55	0.016 0.041	55	0.033 0.106
	Experimental	52	0.062 0.121	49	0.064 0.135
Norm (N)	Control	55	0.002 0.015	55	0.005 0.034
	Experimental	52	0.015 0.111	49	0.000 0.000

Table 9 ANOVA Results for Comparisons by Experimental Condition and Paper Strength

Source	**Sum of Squares**	***df***	**Mean Square**	***F***	***p***	**Partial η^2**	**Obsrvd. Power**
Comparison to a Stipulated Criterion (C)							
Omnibus Model	*0.008*	*3*	*0.003*	*1.192*	*0.902*	*0.003*	*0.086*
Condition (Control vs. Experi.)	0.005	1	0.005	0.317	0.574	0.002	0.087

Source	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial η^2	Obsrvd. Power
Paper (Lower vs. Higher)	0.002	1	0.002	0.144	0.705	0.001	0.066
Condition x Paper	0.002	1	0.002	0.140	0.709	0.001	0.066
Comparison to the Student's Own Other Work (S)							
Omnibus Model	*0.084*	*3*	*0.028*	*2.504*	*0.060*	*0.035*	*0.614*
Condition (Control vs. Experi.)	0.076	1	0.076	6.788	0.010	0.032	0.737
Paper (Lower vs. Higher)	0.005	1	0.005	0.413	0.521	0.002	0.098
Condition x Paper	0.003	1	0.003	0.284	0.595	0.001	0.083
Comparison to the Norm of Other Students' Work (N)							
Omnibus Model	*0.007*	*3*	*0.002*	*0.716*	*0.543*	*0.010*	*0.201*
Condition (Control vs. Experi.)	0.001	1	0.001	0.302	0.583	0.001	0.085
Paper (Lower vs. Higher)	0.002	1	0.002	.0642	0.424	0.003	0.125
Condition x Paper	0.004	1	0.004	1.246	0.266	0.006	0.199

As predicted, no main effects appeared for comparisons to the stipulated criteria (C) or to the norm of other students' work (N). Also as predicted, a small main effect appeared for self-comparative comments (S) as related to experimental condition, $F(1, 207) = 6.788$, $p = 0.010$, partial $\eta^2 = 0.032$. No main effects were present for weaker versus stronger papers, nor were there any significant interactions between experimental condition and paper strength.

These results suggest the partial support of $H3_B$. As predicted, comparisons to the stipulated for the criteria for "good writing" remained constant across experimental conditions and relative paper strength. So too did comparisons to the norm of other students' work. Moreover, as predicted, teachers under the experimental condition offered higher rates of self-comparative feedback (0.063, SD 0.127) than did their control-group colleagues (0.025, SD 0.081). Predicted gains in self-reflective feedback for weaker versus stronger papers, however, were not present.

$H3_C$: The proportional amounts of descriptive and evaluative comments will remain

constant across experimental conditions and relative paper strengths, as teachers' responses are likely to be similarly descriptive or evaluative regardless of whether they are responding to the student's text per se or to the student's comments about that text.

Table 10 provides descriptive sample sizes, means, and standard deviations for experimental assignment, paper strength, and feedback cohort, with respect to the comments' *descriptive* (D) or *evaluative* (E) functions. Table 11 provides summary information for the 2x2 ANOVA on the main and interaction effects for these elements of function within teachers' feedback.

Table 10 Descriptive Statistics for Function by Experimental Condition and Paper Strength

	Condition	***Weaker***		***Stronger***	
		n	Proportional Mean, SD	n	Proportional Mean, SD
Descript. (D)	Control	55	0.088 0.190	55	0.789 0.243
	Experimental	52	0.869 0.177	49	0.814 0.238
Eval. (E)	Control	55	0.121 0.190	55	0.213 0.242
	Experimental	52	0.131 0.177	49	0.186 0.238

Table 11 ANOVA Results for Feedback Function by Experimental Condition and Paper Strength

Source	**Sum of Squares**	***df***	**Mean Square**	***F***	***p***	**Partial η^2**	**Obsrvd. Power**
Descriptive Comments (D)							
Omnibus Model	*0.306*	*3*	*0.102*	*2.235*	*0.085*	*0.031*	*0.560*
Condition (Control vs. Experi.)	0.003	1	0.003	0.062	0.804	< 0.001	0.057
Paper (Lower vs. Higher)	0.280	1	0.280	6.133	0.014	0.029	0.693
Condition x Paper	0.016	1	0.016	0.356	0.551	0.002	0.091

Source	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial η^2	Obsrvd. Power
Evaluative Comments (E)							
Omnibus Model	*0.312*	*3*	*0.104*	*2.286*	*0.080*	*0.032*	*0.571*
Condition (Control vs. Experi.)	0.003	1	0.003	0.072	0.789	< 0.001	0.058
Paper (Lower vs. Higher)	0.284	1	0.284	6.244	0.013	0.029	0.701
Condition x Paper	0.017	1	0.017	0.381	0.538	0.002	0.094

Small but significant main effects appeared for D and E with respect to paper quality. Weaker papers received a greater proportion of descriptive comments (0.874, SD 0.183) than did stronger ones (0.800, SD 0.239). While this difference was significant, $F(1, 209) = 6.133$, $p = 0.014$, the effect size was small, partial $\eta^2 = 0.029$. By the same token, stronger papers received a greater proportion of evaluative comments (0.200, SD 0.239) than did weaker ones (0.126, SD 0.183). Again the difference was significant, but the effect size small, $F(1, 209) = 6.244$, $p = 0.013$, partial $\eta^2 = 0.029$.

These results suggest support for $H3_C$ with respect to its tenet on experimental condition, as teachers in the control and experimental conditions offered roughly the same proportions of descriptive (0.834, SD 0.222; 0.842, SD 0.210) and evaluative feedback (0.167, SD 0.221; 0.158, SD 0.209). However, the paper strength-related portion of $H3_C$ was broken, with stronger papers receiving a greater proportion of evaluative comments (0.200, SD 0.239) than did weaker ones (0.126, SD 0.183).

$H3_D$: A higher proportion of comments will possess positive valence in the experimental condition and with higher-quality papers, as teachers in both situations will adopt a model of communication best described as evaluator-to-person rather than evaluator-to-text. This is to say that as teachers respond to better papers and to papers supplemented with student-provided commentaries under the experimental condition, they will more frequently rise above mere

valence-neutral language of editorial symbols and simple edits, and into domains of communication that involve a more interpersonally "positive" and engaging manner of describing the text's strengths and weakness.

Table 12 provides descriptive sample sizes, means, and standard deviations for experimental assignment, paper strength, and feedback cohort, with respect to the comments' *valence,* either *positive* (🡅) or *negative* (🡇). Table 13 provides summary information for the 2x2 ANOVA on the main and interaction effects for these elements of valence within teachers' feedback.

Table 12 Descriptive Statistics for Valence by Experimental Condition and Paper Strength

	Condition	***Weaker***		***Stronger***	
		n	Proportional Mean, SD	n	Proportional Mean, SD
Positive (🡅)	Control	55	0.187 0.213	55	0.297 0.308
	Experimental	52	0.269 0.237	49	0.326 0.233
Negative (🡇)	Control	55	0.162 0.244	55	0.106 0.180
	Experimental	52	0.112 0.219	49	0.105 0.139

Table 13 ANOVA Results for Valence by Experimental Condition and Paper Strength

Source	**Sum of Squares**	***df***	**Mean Square**	***F***	***p***	**Partial η^2**	**Obsrvd. Power**
Positive Comments (🡅)							
Omnibus Model	*0.575*	*3*	*0.192*	*3.038*	*0.030*	*0.042*	*0.708*
Condition (Control vs. Experi.)	0.163	1	0.163	2.578	0.110	0.012	0.359
Paper (Lower vs. Higher)	0.367	1	0.367	5.820	0.017	0.027	0.671
Condition x Paper	0.039	1	0.039	0.617	0.433	0.003	0.122

Source	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial η^2	Obsrvd. Power
Negative Comments (🡻)							
Omnibus Model	*0.122*	*3*	*0.041*	*1.009*	*0.390*	*0.014*	*0.272*
Condition (Control vs. Experi.)	0.034	1	0.034	0.847	0.359	0.004	0.150
Paper (Lower vs. Higher)	0.054	1	0.054	1.331	0.250	0.006	0.209
Condition x Paper	0.031	1	0.031	0.765	0.383	0.004	0.140

For positive-valence comments, a significant, small main effect appeared across the variable of paper quality, $F(1, 209) = 5.820$, $p = 0.017$, partial $\eta^2 = 0.027$. No other main affects appeared.

These results provide mixed support for $H3_D$. No effects were found according to experimental condition. Stronger papers, however, did receive a small but appreciable increased proportion of positive-valence comments (0.311, SD 0.274) than weaker papers (0.227, SD 0.228). Statistically, this difference registered as small (partial $\eta^2 = 0.027$), but practically speaking, stronger papers received about 1 ⅓ more of such comments.

$H3_E$: The proportions of comments judged to be "clear" or "unclear" will remain constant across conditions and degrees of paper strength.

Table 14 provides descriptive sample sizes, means, and standard deviations for experimental assignment and paper strength, with respect to the comments' clarity, judged either as *clear* (✓) or *unclear* (?). Table 15 provides summary information for the 2x2 ANOVA on the main and interaction effects for clarity within teachers' feedback.

Table 14 Descriptive Statistics for Clarity by Experimental Condition and Paper Strength

	Condition	*Weaker*		*Stronger*	
		n	Proportional Mean, SD	n	Proportional Mean, SD
Clear (✓)	Control	55	0.847 0.176	55	0.826 0.296
	Experimental	52	0.837 0.206	49	0.854 0.171
Unclear (?)	Control	55	0.150 0.172	55	0.159 0.263
	Experimental	52	0.147 0.177	49	0.144 0.171

Table 15 ANOVA Results for Feedback Clarity by Experimental Condition and Paper Strength

Source	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial η^2	Obsrvd. Power
Clear Comments (✓)							
Omnibus Model	*0.024*	*3*	*0.008*	*0.163*	*0.9221*	*0.002*	*0.080*
Condition (Control vs. Experi.)	0.005	1	0.005	0.098	0.754	< 0.001	0.061
Paper (Lower vs. Higher)	< 0.001	1	< 0.001	0.005	0.943	< 0.001	0.051
Condition x Paper	0.019	1	0.019	0.385	0.535	0.002	0.095
Unclear Comments (?)							
Omnibus Model	*0.007*	*3*	*0.002*	*0.057*	*0.982*	*0.001*	*0.060*
Condition (Control vs. Experi.)	0.004	1	0.004	0.108	0.743	0.001	0.062
Paper (Lower vs. Higher)	0.001	1	0.001	0.013	0.911	< 0.001	0.051
Condition x Paper	0.002	1	0.002	0.050	0.823	< 0.001	0.056

No main effects or interactions were observed with respect to clarity, lending support toward the confirmation of $H3_E$.

$H3_F$: The proportions of comments judged to be "specific" or "unspecific" will remain constant across conditions and degrees of paper strength.

Table 16 provides descriptive sample sizes, means, and standard deviations for experimental assignment and paper strength, with respect to the comments' specificity, judged either as *appropriately specific* (A) or *not appropriately specific* (~~A~~). Table 17 provides summary information for the 2x2 ANOVA on the main and interaction effects for specificity within teachers' feedback.

Table 16 Descriptive Statistics for Specificity by Experimental Condition and Paper Strength

	Condition	***Weaker***		***Stronger***	
		n	Proportional Mean, SD	n	Proportional Mean, SD
Specific (A)	Control	55	0.713 0.298	55	0.698 0.331
	Experimental	52	0.764 0.241	49	0.763 0.258
Not Spcf. (~~A~~)	Control	55	0.279 0.283	55	0.299 0.331
	Experimental	52	0.236 0.241	49	0.234 0.256

Table 17 ANOVA Results for Feedback Specificity by Experimental Condition and Paper Strength

Source	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial η^2	Obsrvd. Power
Appropriately Specific Comments (A)							
Omnibus Model	*0.185*	*3*	*0.062*	*0.758*	*0.519*	*0.011*	*0.211*
Condition (Control vs. Experi.)	0.179	1	0.179	2.205	0.139	0.011	0.315
Paper (Lower vs. Higher)	0.003	1	0.003	0.037	0.847	< 0.001	0.054
Condition x Paper	0.002	1	0.002	0.028	0.868	< 0.001	0.053
Not Appropriately Specific Comments (~~A~~)							
Omnibus Model	*0.165*	*3*	*0.055*	*0.698*	*0.554*	*0.010*	*0.197*
Condition (Control vs. Experi.)	0.154	1	0.154	1.950	0.164	0.009	0.285
Paper (Lower vs. Higher)	0.005	1	0.005	0.061	0.806	< 0.001	0.057
Condition x Paper	0.006	1	0.006	0.079	0.779	< 0.001	0.059

No main effects or interactions were observed with respect to specificity, lending support toward the confirmation of $H3_F$.

$H3_G$: As with valence—a measure of "positive" communication, even when communicating the necessary improvements to a text—the proportion of comments judged to be helpful in tone (respectful, positioning the student as agent) will be greater in the experimental condition and with stronger papers.

Table 18 provides descriptive sample sizes, means, and standard deviations for experimental assignment and paper strength, with respect to the comments' helpfulness, judged either as *helpful* (H) or *unhelpful* (~~H~~). Table 19 provides summary information for the 2x2 ANOVA on the main and interaction effects for helpfulness within teachers' feedback.

Table 18 Descriptive Statistics for Helpfulness by Experimental Condition and Paper Strength

	Condition	*Weaker*		*Stronger*	
		n	Proportional Mean, SD	n	Proportional Mean, SD
Helpful (H)	Control	55	0.206 0.231	55	0.317 0.312
	Experimental	52	0.269 0.237	49	0.332 0.231
Not Hlpfl. (~~H~~)	Control	55	0.158 0.244	55	0.100 0.183
	Experimental	52	0.105 0.203	49	0.099 0.138

Table 19 ANOVA Results for Helpfulness by Experimental Condition and Paper Strength

Source	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial η^2	Obsrvd. Power
Helpful Comments (H)							
Omnibus Model	*0.519*	*3*	*0.173*	*2.640*	*0.051*	*0.037*	*0.640*
Condition (Control vs. Experi.)	0.080	1	0.080	1.214	0.272	0.006	0.195
Paper (Lower vs. Higher)	0.400	1	0.400	6.106	0.014	0.029	0.691
Condition x Paper	0.032	1	0.032	0.493	0.483	0.002	0.108
Not Helpful Comments (~~H~~)							
Omnibus Model	*0.132*	*3*	*0.044*	*1.134*	*0.336*	*0.016*	*0.303*
Condition (Control vs. Experi.)	0.039	1	0.039	1.017	0.314	0.005	0.171
Paper (Lower vs. Higher)	0.054	1	0.054	1.387	0.240	0.007	0.216
Condition x Paper	0.035	1	0.035	0.904	0.343	0.004	0.157

Main effects were present for *helpful* (H) comments with respect to paper quality but not to condition. Paper strength produced a small, significant effect, $F(1, 209) = 6.106$, $p = 0.014$, partial $\eta^2 = 0.029$.

Altogether, these results provide mixed support for $H3_G$. While the experimental

condition itself failed to support the hypothesis of better tonal helpfulness, paper strength did account for an appreciable effect, with stronger papers receiving a greater proportion of helpful comments (0.324, SD. 0276) than did weaker ones (0.236, SD 0.235).

Summary of Findings

This chapter sought to answer the eight research questions posed in chapter one. Based on the collected data, the study's results are mixed but generally promising.

This was true, in the first case, with respect to matters of average scores and interrater reliability. Contrary to an undesired expectation, the self-feedback routine did not precipitate a meaningful change in average grades in three of the four tested conditions. And even in the fourth condition (*HSB*, the weaker high school paper), the inflation was less than a full letter grade, raising the group average from 71.9 (SD 9.139) to a 78.0 (SD 6.110). Concomitantly, intraclass correlations improved among middle school teachers. As measured by the intraclass correlation, their absolute agreement rose from r = 0.679 to r = .862. And although high school teachers did not see a similar improvement—their experimental condition actually seeing a *decrease* in absolute agreement to r = 0.960 from the control-group's absolute agreement of r = 0.978—their group-wise IRR was nevertheless strong by traditional standards.

The study's results were mixed with respect to questions about teachers' feedback practices. To begin, the introduction of a student's own self-reflective comments prior to the teacher's grading and feedback cycle did in fact precipitate an at-times beneficial difference across experimental conditions.

- FT and FS remained constant across experimental conditions. And although FP broke its portion of the hypothesis by also remaining constant, the proportion of comments focused on students' self-regulation improved by the magnitude of a

medium-sized effect (partial $\eta^2 = 0.079$).

- Comparisons to stated criteria and to the norm of other students' work remained proportionally constant while comparisons to the same student's imagined previous and/or successive attempts increased as predicted, though only by a small magnitude (partial $\eta^2 = 0.032$).
- The proportion of descriptive versus evaluative comments remained stable across experimental conditions, as predicted.
- Clarity and specificity remained constant, as predicted.
- However, the proportion of positive-valence comments did not increase as predicted across conditions, nor did the proportion of comments employing a helpful tone.

Unfortunately, best-practice feedback was not given consistently across the levels of paper strength. Stronger papers did, as previous studies have implied, receive more holistically generous feedback than weaker ones; but the introduction of student-authored comments did not meaningfully close the gap in quality between the feedback given to weaker versus stronger papers.

- Stronger papers received a greater proportion of evaluative comments than did weaker papers (partial $\eta^2 = 0.034$).
- Stronger papers received a greater proportion of positive-valence comments than did weaker papers (partial $\eta^2 = 0.033$).
- Clarity and specificity remained constant
- Stronger papers received a greater proportion of helpful comments than did weaker ones (partial $\eta^2 = 0.036$).

The above are, truly, summary data—abstract and dry, even to me as the primary researcher. In the following chapter, I will try to make concrete a few various aspects of the underlying realities, particularly with respect to what was to be gained by examining the qualitative aspects of the feedback itself. With even a few examples it may be possible for this project to rise above the mere tallying of results, into the clearer air of something resembling a useful exercise in professional development worthy of the name.

CHAPTER FIVE: DISCUSSION

Introduction

Time. Reliability. Relevance. This study has responded three perennial problems facing classroom practitioners, perhaps best re-introduced thus: *Is there anything we teachers-as-evaluators can do to (a) shift the demands on our time away from tasks that are primarily sorting oriented and toward those that have a higher probability of encouraging student growth, while (b) not trading away our sense of "rigor"—whatever that means—or (c) decreasing our agreement about the grades students have earned?*

The data seem to have responded with a mild "yes" regarding the study's hypotheses. When teachers encounter student texts upon which the students themselves have already placed corrections and/or comments of their own, they are no less likely to remain consistent in the grading, while they are mildly yet significantly more likely to respond with a greater proportion of feedback of the sort proven by previous research to be optimally constructed so as to encourage student growth.

Building on the previous chapter's foundation of quantitative results, then, this final chapter seeks first to discuss the study's rather straightforward data with respect to the questions of rigor and agreement in grading, after which it will pursue a longer discussion regarding two issues central to the study's focus. The first of these involves teachers' use of time, and their reflections on the importance of time as a contextual hedge around the work they do. The second central issue involves the rather complex results pertaining to teacher feedback under the studied conditions, including the role of a meaningful variable not anticipated in the study's original hypotheses—teachers' responses as reflecting their membership in middle- or high school faculties. In part this latter discussion will serve to round out the picture of the quantitative data

themselves, but in part, too, it will turn from the consideration of generalized, numerically rendered outcomes so as to examine the concrete qualitative data from which these generalizations have been drawn—i.e., the teachers' actual feedback. It is hoped that this mixed-methods approach will provide useful insights for fellow practitioners and researchers alike.

Grade Inflation and Reliability

Among the concerns that drove this study was the sense that students perceive a lack of consistency in the grades they receive: some teachers are "easy" graders while others are far too demanding. In my own experience as a college student, reliable voices warned me away from one or two overly aggressive graders in the English department. As a eleven-year teaching veteran, I fall into at least one conversation per semester with other faculty members or with our administrators about students who feel their grades are reflective more of a teacher's idiosyncratic vision of "good writing" than of the paper's actual merits. Sometimes these students are adamant enough in expressing their frustration—and convincing enough in their arguments—that they are granted the relatively rare luxury of transferring into another teacher's section. Similarly, whenever the ELA teachers of my district have convened to score the Kansas Writing Assessment, it has seemed readily apparent some table groups are much more likely than others to depend on third readings to arbitrate disagreements in scores. Last year, in response to situations such as these, my own principal asked the school's leadership team to read O'Connor's *A Repair Kit for Grading: 15 Fixes for Broken Grades* (2007), a book whose major premise is that the educational world needs to improve our grading practices across each of four domains. According to O'Connor, educators must work toward making our grades more consistent, accurate, meaningful to students, and supportive to the learning process. O'Connor believes this problem looms so large on the academic landscape that he opens the first chapter

with the following challenge from Marzano (2000): “Why [w]ould anyone want to change current grading practices? The answer is quite simple: grades are so imprecise that they are almost meaningless” (O’Connor, p. 3).

Given this sort of background, I expected to see wild divergence in the scoring habits of the host district’s teachers, with respect both to their reliability and also to the ways they would respond to the hypothetical students’ predicted grades. I was pleasantly surprised to see that there was more agreement among the host district’s ELA teachers than presumed. This agreement demonstrated itself first in the comparative lack of wobble in the average scores given across control and experimental conditions. At the middle school level, the wobble was almost nonexistent. Facing the weaker paper, middle school teachers in both conditions agreed that it merited a 79%. Even the standard deviations for these groups were reasonably close, at 6.276 for the control group and 6.689 for the experimental. Much the same could be said for middle school teachers as they evaluated the stronger paper. Where the control group teachers gave the paper an average score of 82.3 %, those in the experimental group were in the same low-to-mid *B* range with an 83.9%; standard deviations were 6.798 and 6.997 respectively. In both cases, teachers’ close agreement with each other stood in contradistinction to the students’ predicted scores. “John Cauthron,” author of the weaker paper, predicted he should receive a *B*; “Roger Hengst,” author of the stronger paper believed his work had merited an *A*.

At the high school level, the story was somewhat—but not entirely—different, with the major differences pertaining to the weaker paper. On this paper, written by “Paula Healey”—who predicted she would receive a “low *B*”—experimental group teachers in fact believed the paper merited a 78.0% (SD 6.110) while control group teachers scored the paper lower, at a 71.88% average (SD 9.139). Admittedly, the group sizes for this analysis are on the small side

(N = 25 for each group) and therefore susceptible to the effects of outliers (skew for the control group was -0.530, for the experimental 0.091). For this reason it is worth noting that the mode scores for the control group were 70% and 75% (6 teachers at each level), while the mode for the experimental group was 80%. Experimental-group teachers really do seem to have been affected by the students' suggested score, at the magnitude of just over one-half a letter grade. Yet at the same time, their scores are less widely distributed than are those from the control group; so while they were more forgiving than their control-group counterparts, they enjoyed a higher degree of agreement about the scores they assigned. By contrast, when facing "Samantha Miller's" stronger paper, high school teachers in both groups reached similar conclusions about its merits. Control-group teachers awarded an average score of 86.8% (SD 6.276), a score minimally higher than the experimental group's 86.2% (SD 5.520). The experimental group seems not to have been meaningfully impacted in this case by Miller's own predicted score of a "Low A/High B."

Thus, experimental-group teachers in three of four trials did not typically succumb to what I predicted would be a practically relevant degree of grade inflation; and for this I am quite glad, as these results seem to open a door for a friendlier, more socially proactive pedagogical practice without the barrier of fear that "friendly" must equate to "soft." I am also happy to report that the study demonstrated a profoundly positive impact on agreement in scores where it was most needed, among the host district's middle school ELA faculty. By the ICC measure of absolute agreement—perhaps the most meaningful lens from a student's point of view—the control group of these middle school teachers achieved only r = 0.679 for group-wide reliability, r = 0.066 for the reliability of a single rating. While the group-wide number would be low enough to be problematic for high-stakes testing purposes, the single-rating agreement is very slight indeed. Under the experimental condition, the picture improved somewhat for single

ratings (r = 0.178) and strongly enough for group-wide considerations (r = 0.862) to meet the 0.80 standard established by Diederich and others.

With high school teachers the group-wide reliability was quite strong for both the control (r = 0.978) and experimental (r = 0.960) conditions. Unfortunately, the experimental condition weakened ICC single-rating results from the control's r = 0.641 to r = 0.491[18]. But that having been said, were I a high school student in this district, I might nevertheless be inclined to submit a self-annotated essay, as in this dataset the lowered reliability value traveled with a half-letter average grade increase relative to the outcomes for unadorned final copies.

The host district's teachers themselves offered only few unsolicited comments about their grading practices or the scores offered to the study's essays. One middle school teacher balked at the notion of a percentage grade, writing, "I don't like giving a score in 8th [grade]; it's either *A*, *B*, *C*, or *D* quality" (ME13a). Another reflected that her grade of 85% on the weaker of the two middle school papers was "probably too high" because she "rushed" in completing the task (MC14b). No comments whatsoever gave any impression that teachers in the experimental condition felt their grades had been skewed by the students' predicted scores; and in most cases, this impression is borne out by the evidence—the one exception being high school teachers when grading the weaker paper.

Time

Time: Admittedly, the research task itself involved a contrived scenario, but one for which the time constraints were intended to be reasonably authentic to the practice of working

[18] This result is somewhat surprising, given that when the "consistency" lens is substituted for "absolute agreement," ICC single-rating results actually improved from r = 0.681 to r = 0.706.

teachers. With respect to that time frame, this study's six minutes per essay seems at first a too-short interval for the offering of a grade and feedback, especially when approximately two of those minutes must be consumed by the simple task of reading itself. Yet considered against the reality that teachers in the host district might have as many as 125 students on their rosters—requiring from them an additional two hours of evaluation time for each additional minute they give per paper—six minutes may not be far from a realistic balance between economy of time and strength of response. Certainly, many admirable examples of robustly informative and helpful feedback appeared in the dataset.

That said, a handful of teachers—without any specific prompting beyond "Please feel free to write any initial comments you have about the work you just completed"—noted their sense of frustration with the time allotted for their work. Given that only a few teachers made such comments, no generalizations are possible from the information they have provided. Yet their comments are nevertheless interesting and potentially illuminating in at least two ways. First, whenever such comments were made they never suggested that *too much* time had been allowed. Second, there appeared to be a degree of parallelism between the manner and tone of teachers' comments about time constraints and their comments about the papers themselves; and where such parallelism existed, it seemed to provide tentative evidence that teachers' ways of communicating with students about their work may be reflective not simply of their training to look for errors (Murray, 1982), but also of their general ways of communicating about academic concerns. Some of us by nature are cranky, others congenial. Some are quite willing to work cheerfully within a broad range of circumstances, while others are more inclined toward dourness even on Friday afternoons at 3:15. Similarly, some will always place blame for our "failures" on external limitations, while others will mediate the blame by reflecting on the

variables within our own mental sets. What is thus is most interesting about teachers' comments about time is not so much their frustration *per se* but the various means by which they chose to express this frustration.

For example, one middle school teacher's time-related comments were cheerily apologetic as she wrote, "I didn't have enough time to respond—I'm a slow grader. ☺" (ME9a).[19] Here, the tonal qualities are consistent with the teacher's feedback to the essays themselves. For example, on the weaker essay she included two annotations rising above simple edits in such a way as to take on tonal characteristics. The first of these is in response to the hypothetical student's question, "I need details to show why, right?" (Appendix D, ¶1, l. 5): "Can you give me an example of what your mom has done to teach you this lesson? Be specific" [WCD↑✓AH]. A second annotation was a summative remark at the paper's end, expressing an overall positive outlook on the paper while reinforcing the message of approximately five simple edit-level notations throughout the paper: "I think you have a great start, but correcting some grammar issues and adding more specific details will make your paper better. ☺" [WOCSE↑✓AH].

Taken together, these comments demonstrate reasonably well Brookhart's (2008) desirable feedback characteristics. By leading off with an interrogative sentence highlighting the student's success in delivering both a "teacher" and a "lesson," the first comment shows both respect for the author and also the teacher's desire to be understood as an interested reader of the student's ideas, not merely an academic critic. Only after this sense has been established does the

[19] Coding for parentheticals of teacher comments is as follows: M = middle school, H = high school; C = control, E = experimental; # = number within condition; a = stronger paper, b = weaker paper.

teacher then turn to the mildly imperative "Be specific" element. Similarly, the summative comment opens with praise for what the student has already accomplished with the current draft. While greater specificity about the praiseworthy elements would have improved this comment's merits, it has already accomplished much by its friendly tone and the illumination of grammar and details as two areas for further improvement—both of which qualities have received specific attention among the teacher's simple edits. Returning to the notion of teachers' sense of time's role in the task set before them, it comes as no surprise that such comments should come from a teacher who reflected about her work on the weaker paper, "I try to give positives first! If I had more time, I would've responded more" (ME9a).

Other teachers' time-related comments were flatly factual, as with one middle school teacher's "Ran out of Time" notes (MC4a, b) and another's "Did not finish" (MC8a, b). Again, not unlike the cheery teacher above, the flatness of these participants' comments about the research task seem somewhat aligned with their manner of providing feedback to the essays' hypothetical students. Participant MC8, for example, wrote an average of 22.5 comments per essay, just over twice the overall average of 10.47 comments; yet on the weaker of the two papers s, every one of her 18 comments represents a simple edit—usage, tense shifts, pronoun reference, comma errors, and the like. There is no conversation whatsoever with the student behind the paper. In fact, a total of 10 words have been written on the page, four of which have been incorporated into the comment set's longest statement "Don't end with it" [WCD✓A] (MC8b). Similarly, 25 of the stronger paper's 27 comments are simple edits, while the comments receiving codes for valence and tone are so flatly stated that an argument could be made for their

miscoding.[20] Both of these comments appear to the left of the paper's first paragraph [Appendix A]. Toward the paragraph's beginning is a comment drawing attention to two essential criteria for an essay introduction: "1st: Intro.—attract reader, state purpose" [WCD🡻✓A~~H~~]. Just below it appears a second comment directing the student again toward stated criteria: "Too long of an introduction. Make two paragraphs" [WCD🡻✓~~AH~~]. In both cases the comments have been coded as displaying negative valence and tone, calling attention to problems within the paper but not offering concrete suggestions as to their remedy and, therefore, not portraying helpfulness so much as bossiness to the student. In fact, participant-teacher MC8 herself revealed self-awareness of the comments' negativity not only in her reflections that "I was too nitpicky" and "I was not positive enough!" but also in the revised comments she added during the professional development session following the study's data collection sequence, among which are "Good use of fig. language" and "Great beginning" (MC8a). And while these comments still miss the mark with respect to the criterion of specificity, they nevertheless represent a tonal step in the right direction.

Participant-teacher MC4 demonstrated somewhat similar flatness in her comments, again

---

[20] As we worked through the coding process, both coders occasionally found ourselves untangling the valence/tone knot by asking ourselves how "simple" a simple edit might seem to the four essays' hypothetical authors (stronger, weaker; eighth- and twelfth-grade), rather than to ourselves as coders or to what we imagined might be the participant-teachers' intentions. And as we considered comments and codes through this lens, we gained a new layer of appreciation in our understanding of reader-response theory. To a professional editor or a graduate of advanced literary studies, for example, the comment "Too long of an introduction. Make two paragraphs" might seem entirely straightforward, hence *simple,* edit. But to the eighth-grade author of "What I Leanred from My Dad," the suggestion might feel like an overwhelming obstacle.

in parallel with her flatly factual comments about time's role in her morning's work. 10 of her 15 annotations on the weaker paper are simple edits, while three of the remaining five fail reach meaningful success with respect to valence, specificity, or helpfulness:

- "Grammatical errors that get in the way of meaning." [WCD🡻✓~~AH~~]
- "Lots of repetition, but drove points home." [WCD🡹✓~~A~~H]
- "Tried to use commas, but sometimes incorrectly." [WCD🡻✓~~A~~H]
- "Interesting anecdote/story." [WCE🡹✓AH][21]
- "Conclusion was full circle/Nice job!" [WCE🡹✓AH]

Among these, only the "Interesting anecdote/story" and "Nice job!" annotations seem to ring with notable enthusiasm. And while both received codes for adequate specificity, neither pushes beyond a bare minimum toward illuminating what exactly is interesting or nice about the student's work.

Perhaps what is most remarkable with this comment set, however, is the contrast provided by the teacher's supplementary annotations provided during the professional development session following the initial data-collection task:

- "You really have some great goals in your life" (MC4b).
- "Your grammar is on the way" (MC4b).
- "I'm proud of you! Great voice & word choice!" (MC4a).

---

[21] The *A* code here is arguable. While it would have been nice if something specific within the anecdote had been highlighted as providing interest, we agreed that the comment's proximity to paragraph three [Appendix C] provided enough of a contextual lever to count for specificity. A similar caveat could be offered for MC4's final comment.

- "Lots of anecdote!!" (MC4a).
- "I can tell you're working on conjunctions" " (MC4a).
- "Very creative" (MC4a).

While all six of these revisions would benefit from a greater incorporation of details to support their general ideas, they nevertheless seem to reflect the teacher's sense that her original comments are not likely to be seen as emotionally engaging, warm, or inviting. Thus, even without providing self-reflective comments like MC8's, MC4 seems to have achieved at least one short-term insight from exposure to Brookhart's principles in good feedback. Were the time allowed for long-term professional development in evaluative feedback, it seems reasonable that MC4's interest in providing tonally helpful comments might be leveraged into further gains across the feedback spectrum.

Not all comments about time fit in the "cheery" to "flat" ranges of the emotional spectrum. Another teacher veered into a negatively worded frustration with time's pressure. "NOT DONE. Need more time to grade," wrote this middle school teacher (MC28b); "I did not have enough time to go through the standards. I would like to have given him some comments on his work." Similarly to MC8 above, this teacher gave 21 comments overall to the lower-quality essay, only 3 of which rose out of the "simple edit" category, as though it really were necessary to take care of all the "errors" before speaking to the strengths. As might be expected from the comment about time, the general tenor of these 3 comments is somewhat abrasive:

- "Very repetitive" [WCD🡻✓A~~H~~].
- "Very wordy & not clear" [WCD🡻✓~~AH~~].
- "You need to clearly organize your thoughts. Try using a graphic organizer before you start" [WORCD🡹✓AH].

Interestingly enough, MC28 wasn't as limited in her tonal register with the stronger paper, where several of her comments prove she is capable of offering pedagogically rich feedback, as with the following comments:

- "Organization—Good, but maybe don't repeat the clang swoosh. Can you come up with another onomatopoeia?" [WCE↑✓AH].
- "Content organization—You are almost there. I think if you read your essay aloud you will see where a little more organization would be helpful" [WORCSD↑✓AH].

Given that this essay, too, received the annotation "NOT DONE!" (MC28a), the disparities in valence and tone are intriguing and several explanations are no doubt plausible. The simplest might lie in the teacher's own comment that "I think I had a better idea of how I wanted to grade [the better paper]" (MC28a). Another might be related to Chase's (1983) hypothesis that evaluators penalize disproportionately as their frustration levels go up, so that her frustration with the weaker paper interfered with the ability to find ready-at-hand a few positive characteristics to compliment.

Regardless of the cause for her sometimes abrasive tone, MC28—much like MC4 with her "Ran out of time" comment—places the blame for what she perceives to be incomplete work on the constraint of time. "Need more time to grade," she writes, because "I would like to have given him some comments on his work" (MC28b). With both teachers, moreover, the blame is largely external. The teacher *would* have done so much more, if not for the ending bell.[22]

[22] The solitary high school teacher who provided any sort of comment about time echoed this externalized version of the problem's root: "[My comments were] general, but if I had more time, I'd add quite a bit more"

Another middle school teacher, however, saw the matter somewhat differently. This participant blamed her shortage on time neither on the task itself nor the context in which it was to be completed, but on her own way of going about the work of evaluation and feedback. "I ran out of time with this one because I was so focused on errors," she wrote on the weaker paper; "I didn't give a lot of helpful feedback" (MC14b). MC14's comment contains the seeds of what for me was the impressive lesson to be drawn from a qualitative consideration of this study's dataset: *Benefits from the quality of teachers' comments far outweigh the benefits to be gained from the quantity of comments.* It is to this lesson I will soon turn after a brief discussion of one surprising outcome in the dataset—the difference in feedback attributable not to paper quality or experimental condition but to whether the teacher worked with middle- or high school students.

Strength of Feedback—Feedback across the Reconsidered 2x2x2 Conditions

Perhaps the most important intended element of this study was its focus on how the experimental condition would affect teachers' feedback practices. Given students' own comments about the works they had authored, would teachers be more likely to engage in feedback practices shown in research to be supportive of the learning process? The answer seems to be a guarded, limited *yes*. Teachers in the experimental condition were more likely to provide comments focusing on students' self-regulation (partial $\eta^2 = 0.087$) and comparing the present

---

(HE22a). While again stressing that the frequency of comments is far too low to make any meaningful generalizations, it was interesting to note the virtual absence of time-related complaints among high school teachers. Taken together with high school teachers' lower rate of average total comments, the comparative richness of their comments, their comparative responsiveness to the experimental condition, and their appreciably higher degree of interrater reliability, it would seem that high school teachers in the host district viewed the study's task differently than did middle school teachers. Perhaps the nature and causes of this difference would be worthy of further study.

draft to (imagined) previous or successive texts by the same student (partial $\eta^2 = 0.042$).

Another factor anticipated to be meaningfully relevant to teachers' feedback practices was the relative strength of the essays themselves. As expected, stronger papers received a small but significantly greater proportion of positive-valence comments (partial $\eta^2 = 0.033$) and tonally helpful comments (partial $\eta^2 = 0.036$) than did weaker papers. These results were not surprising, given Chase's (1983) hypothesis that "any condition that complicates readability should reduce scores" and Marshall's (1967) belief that easy-to-spot errors are subject to disproportionate grading penalties. Reduced scores and the imposition of disproportionate penalties do not travel hand-in-hand with the mindset required for the parceling out of meaningful praise.

While these findings were not surprising, what was almost totally unexpected was the degree to which teachers' memberships in the middle- or high-school cohorts played into the quality of feedback they might provide. You may recall that this study was originally conceived as a project involving only high school teachers. As must no doubt be often the case in educational research, my negotiations with the host district's gate-keepers involved making accommodations to suit their needs and desires as well as my own. One of the district's expressed desires was that the professional development session be offered to the district's entire secondary ELA faculty, not simply its high school subset. As it turned out, what was from my perspective an annoying complication proved to be the source of the study's most interesting findings, namely that middle- and high school teachers respond to student texts in markedly different ways. In fact, the differences between the middle- and high school teachers were at times of even greater magnitude than was the case for either experimental condition or paper strength.

Table 20 accesses the data to show a few easy-to-see ways in which the middle school

cohort's comment sets differed from those provided by high school teachers. An initial difference was that middle school teachers tended to offer a greater number of comments (12.14, SD 6.401) than did high school teachers (9.19, SD 4.950); this difference was significant, $F(1, 209) = 13.696$, $p < 0.001$. Not surprisingly, within this total-comments context middle school teachers also significantly outpaced their high school counterparts in the average number of comments focused on the work itself (12.04, 8.82), providing comparisons to the criteria against the work was to be judged (12.06, 8.86), comments expressed clearly (10.46, 7.84) and with adequate specificity (9.82, 6.78). Once, however, these raw subscale averages are rendered as proportions of the average total comments per cohort, the apparent differences between middle- and high school outcomes for these traits shrink to the point of becoming not particularly noteworthy. For example, 99.18 percent of middle school teachers' comments focused in some way on the essays themselves, as did 95.97% of high school teachers' comments. Similarly, 99.34% of the middle school comments and 96.41% of the high school comments referred in some way to external evaluative criteria. And although the average number of "clear" was lower for both groups, the decrease was again proportionally similar, with 86.16% of middle school teachers' comments deemed "clear" in the coding process, compared to 85.31% of those provided by high school teachers.

Table 20 Quantitative Differences in Selected Feedback Patterns, by Cohort Membership

Category	MS	HS	*F* (1, 209)	Sig.	Partial Eta Squared	Observed Power
Total Comments	12.14	9.19	13.696	< 0.001	0.061	0.958
Comments Focused on the Task	12.04	8.82	15.864	< 0.001	0.071	0.977
Comments Focused on Self-Regulation	0.784	1.056	4.757	0.030	0.022	0.584

Criterion-Referenced Comparisons	12.06	8.86	15.866	< 0.001	0.071	0.978
Self-Referenced Comparisons	0.06	0.63	38.670	< 0.001	0.156	1.000
Comments Bearing Positive or Negative Valence	2.89	3.77	8.021	0.005	0.037	0.805
Clearly Expressed Comments	10.46	7.84	10.635	0.001	0.048	0.901
Comment Expressed with Adequate Specificity	9.82	6.78	14.067	< 0.001	0.063	0.962

What is numerically striking, however, is the degree to which middle- and high school teachers differed in the feedback traits demonstrating an attempt to reach the "whole-person" author behind the text rather than simply the essay itself. In terms of proportions of total comments, middle school teachers were approximately half as likely as high school teachers to discuss matters of student self-regulation (6.46%, 11.59%) or to move beyond simple textual edits into comments exhibiting a personal "voice" with either positive or negative, helpful or dismissive tonal characteristics (23.81%, 41.02%). Moreover, they were much less likely to make imaginative leaps about the hypothetical students' previous or successive drafts of the essays. While 6.86% of high school teachers' comments somehow accessed the imagined future or past of the text at hand, only 0.49% of middle school teachers' comments demonstrated a willingness to look up from the page itself and imagine its context within a student's through-the-course progress as a writer.

It is on account of remarkable differences such as these that the decision was reached to break the omnibus dataset into middle- and high school subsets, thus treating teacher cohort as an independent variable along with placement in the control or experimental condition. Thus, what was originally intended to be a 2x2 factorial ANOVA of teacher commentary habits under

control and experimental conditions with comparatively better and worse papers was reconsidered according to a 2x2x2 analysis, with the following notable results:

- Focus on the Task (FT): With respect to the writing product itself, middle school teachers offered a minimally higher proportion of comments (0.980, SD 0.118) than did high school teachers (0.944, SD 0.126); $F(1, 209) = 4.532$, $p = 0.034$, partial $\eta^2 = 0.022$.
- Focus on the Author's Self-Regulation Processes (FR): By contrast, high school teachers provided proportionally more than twice as much feedback about the authors' self-regulation processes (0.088, SD 0.141) as did middle school teachers (0.041, SD = 0.090); $F(1, 209) = 9.002$, $p = 0.003$, partial $\eta^2 = 0.042$.
- Comparisons to Criteria (C): Middle school teachers gave minimally more frequent criterion-referenced comparisons (0.982, SD 0.117) than did their high school teacher colleagues (0.948, SD 0.120); $F(1, 209) = 4.385$, $p = 0.038$, partial $\eta^2 = 0.021$.
- Comparisons To Imagined Prior/Successive Attempts (S): High school teachers, however, showed a much greater tendency to draw comparisons between the evaluated work and the students' imaginable prior or subsequent drafts (0.085, SD 0.143) than did middle school teachers (0.007, SD 0.033); $F(1, 209) = 35.348$, $p < 0.001$, partial $\eta^2 = 0.148$. Moreover, an interaction between cohort membership and experimental condition showed that while high school teachers responded to the experimental condition by providing even richer feedback according to this trait (0.125, SD 0.161) than did their control-group counterparts (0.047, SD 0.112), middle school teachers in the experimental group (0.008, SD 0.040)

showed no such response providing this feedback characteristic at a rate quite similar to that of the control group (0.006, SD 0.027); $F(1, 209) = 7.921$, $p = 0.005$, partial $\eta^2 = 0.038$. Thus, in both the overall average rate of response and in their responsiveness to the experimental stimulus, high school teachers went about this element of feedback in a dramatically different fashion than did their middle school counterparts, as demonstrated in Figure 17.

- Evaluative (E) versus Descriptive (D) Comments: High school teachers provided proportionally almost twice as many evaluative comments (0.213, SD 0.235) as did middle school teachers (0.119, SD 0.187), $F(1, 209) = 10.300$, $p = 0.002$, partial $\eta^2 = 0.048$. Moreover, there was a small interaction involving cohorts and experimental conditions. As depicted in Figure 18, high school teachers in the control condition were much more likely to offer evaluative feedback (0.257, SD 0.256) than they were in the experimental condition (0.167, SD 0.203) or than middle school teachers were in either the control or experimental conditions (0.150, SD 0.216; 0.917, SD 0.153); $F(1, 209) = 6.729$, $p = 0.010$, partial $\eta^2 = 0.032$.
- Positive-Valence (↑) Comments: Whereas high school teachers tended to offer positive-valence comments about one-third of the time (0.335, SD 0.278), middle school teachers tended toward a greater frequency of "simple edits" (i.e., comments such as spelling and punctuation corrections, for which valence was judged not to come into play), providing positive-valence comments only about one-fifth of the time (0.212, SD 0.219). In other words, high school teachers engaged in positive-valence work 1½ times more frequently than did middle

school teachers, a difference which translates to a significant, medium-sized main effect, $F(1, 209) = 13.138$, $p < 0.001$, partial $\eta^2 = 0.061$.

- Negative-Valence (↓) Comments: Here a small, significant interaction occurred between cohort membership and experimental condition, $F(1, 209) = 4.608$, $p = 0.033$, partial $\eta^2 = 0.022$. As depicted in Figure 19, high school teachers in the control condition (0.192, SD 0.284) expressed negativity much more frequently than they did in the experimental condition (0.103, SD 0.199), or than middle school teachers did in either control or experimental conditions (0.86, SD 0.116; 0.113, SD 0.172). Had this pattern been mirrored in middle school teachers as well, it would have bolstered support of $H3_D$.
- Helpful (H) and Unhelpful (~~H~~) Comments: Middle school teachers gave fewer tonally helpful comments (0.222, SD 0.226) than did high school teachers (0.347, SD 0.280); $F(1, 209) = 13.492$, $p < 0.001$, partial $\eta^2 = 0.062$. As noted already, middle school teachers tended to provide a higher proportion of "simple edits" (1049 of 2384, or 75.8% of their comments) than did high school teachers (522 of 891, or 58.6% of their comments), no doubt accounting for much of this difference in helpfulness. But that having been said, there was no statistically significant difference in the proportion of *unhelpful* comments given with respect to cohort membership, where a smaller overall gap appeared between high school teachers' (0.140, SD 0.243) and middle school teachers' (0.096, SD 0.226) rates of tonally unhelpful comments.

Several interactions were also present. First, cohort membership and paper strength worked to produced interactions both for *helpful* and *unhelpful*

comments. As displayed in Figure 20a, high school teachers were more inclined toward giving comments coded *helpful* with stronger papers (0.431, SD 0.297) than weaker ones (0.339, SD 0.307), or than middle school teachers to either stronger or weaker papers (0.233, SD.0.226; 0.212, SD 0.232); $F(1, 209) = 4.144$, $p < 0.043$, partial $\eta^2 = 0.020$. As displayed in Figure 20b, high school teachers were also more likely to give unhelpful comments to weaker papers (0.191, SD 0.284) than to stronger ones (0.088, SD 0.181), or than middle school teachers to either weaker or stronger papers (0.083, SD 0.146; 0.110, SD 0.145); $F(1, 209) = 5.933$, $p < 0.016$, partial $\eta^2 = 0.028$. Because high school teachers gave helpful and unhelpful comments (i.e., comments that moved beyond simple edits, such as for spelling, and into tonally "active" uses of language) more frequently in general (0.347, SD 0.280; 0.140, SD 0.243) than did middle school teachers (0.222, SD 0.226; 0.096, SD 0.145), that they would outpace middle school teachers' rates for helpful/unhelpful comments is not at all surprising. What is surprising and regrettable is the degree to which high school teachers "rewarded" stronger papers with tonally helpful language while "punishing" weaker papers with tonally unhelpful language.

Second, cohort membership, experimental condition, and paper strength produced an interaction for *helpful* comments. As displayed in Figures 21a and b, high school and middle school teachers provided somewhat unusual response patterns with respect to helpfulness, depending on their assignment to control or experimental groups, as well as whether they were responding to weaker or stronger papers. While middle school teachers in control and experimental

conditions provided proportionally similar rates of helpfulness to weaker (0.212, SD 0.248; 0.212, SD 0.205) and stronger papers (0.182, SD 0.216; 0.291, SD 0.215), high school teachers' helpfulness varied meaningfully across paper strength and experimental conditions; $F(1, 209) = 6.831$, $p < 0.010$, partial $\eta^2 = 0.033$. To stronger papers, their responses were somewhat consistent in the control (0.480, SD 0.335) and experimental groups (0.377, SD 0.245); their helpfulness to weaker papers was somewhat lower in the experimental condition (0.336, SD 0.258), and considerably lower in the control group (0.198, SD 0.197).

Figure 17 Interaction between Cohort & Experimental Condition for Self-Comparison Comments

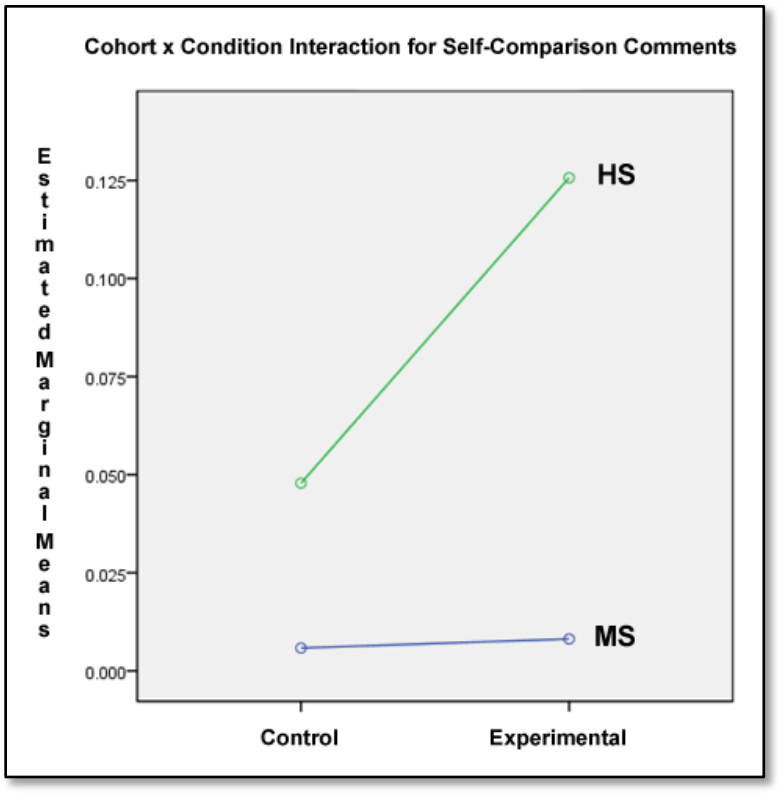


Figure 18 Interaction between Cohort & Experimental Condition for Evaluative Comments

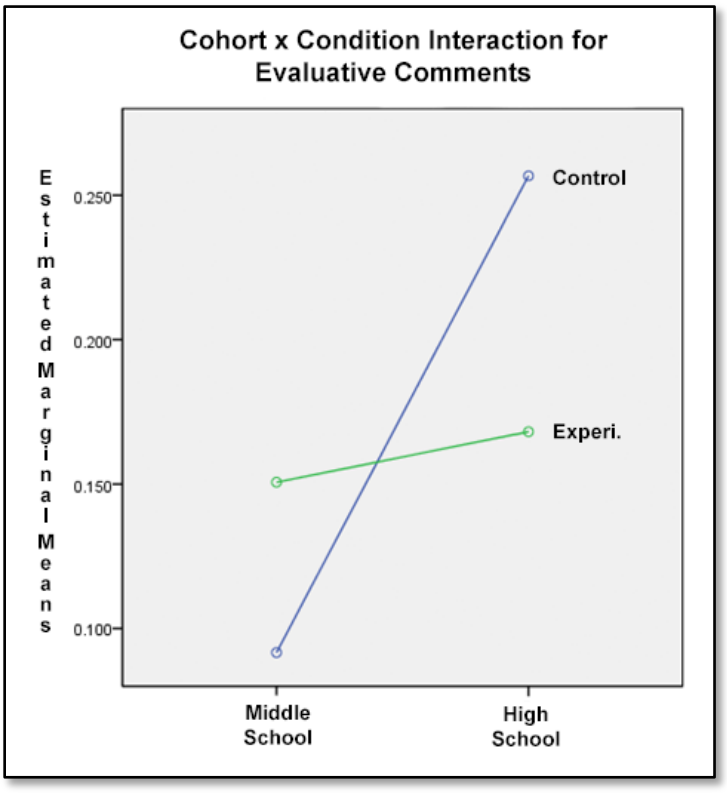
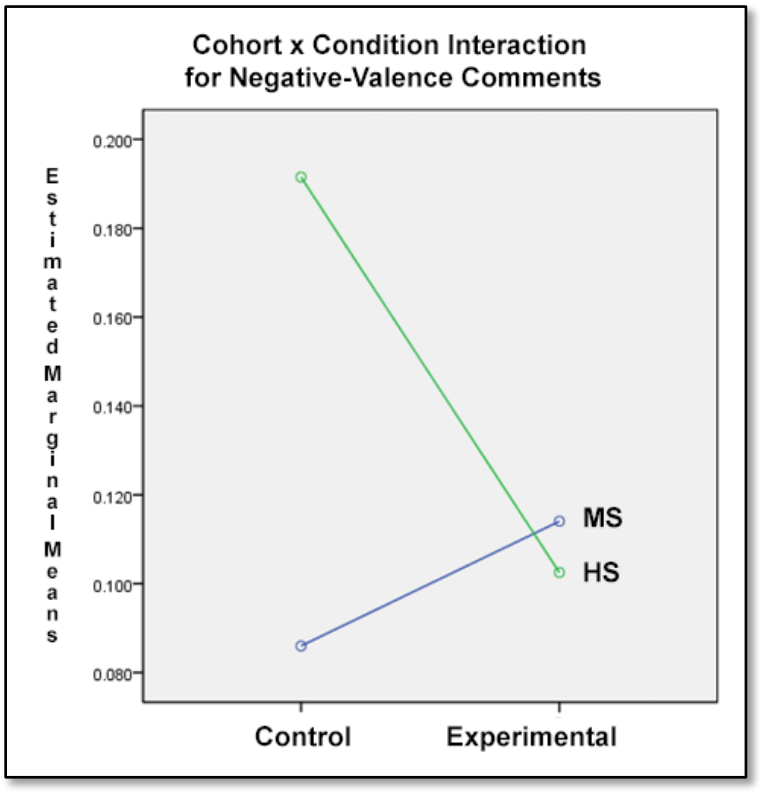
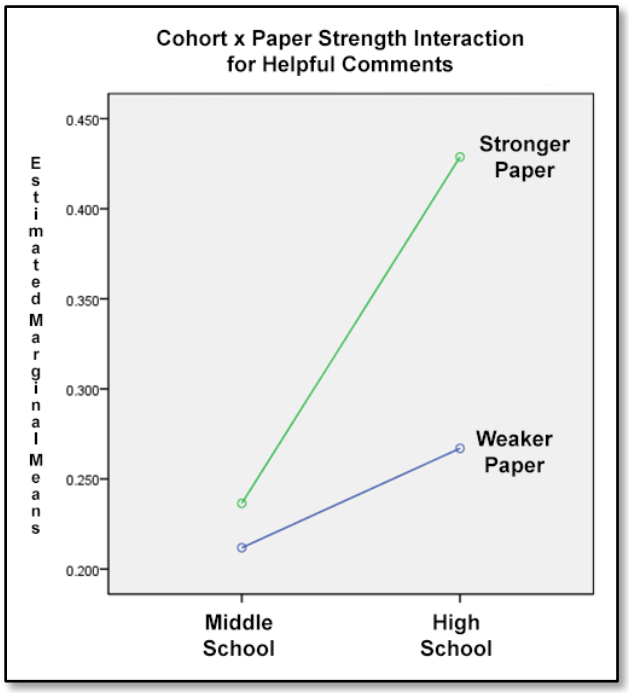
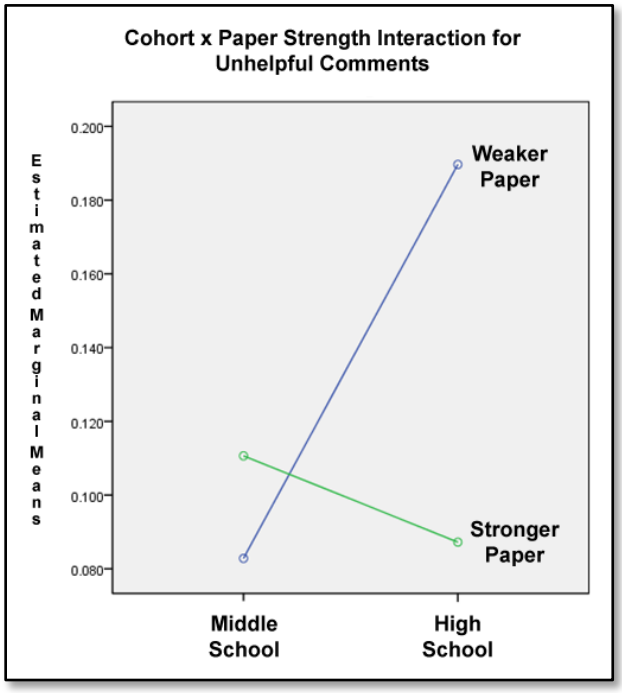


Figure 19 Interaction between Cohort & Experimental Condition for Negative-Valence Comments

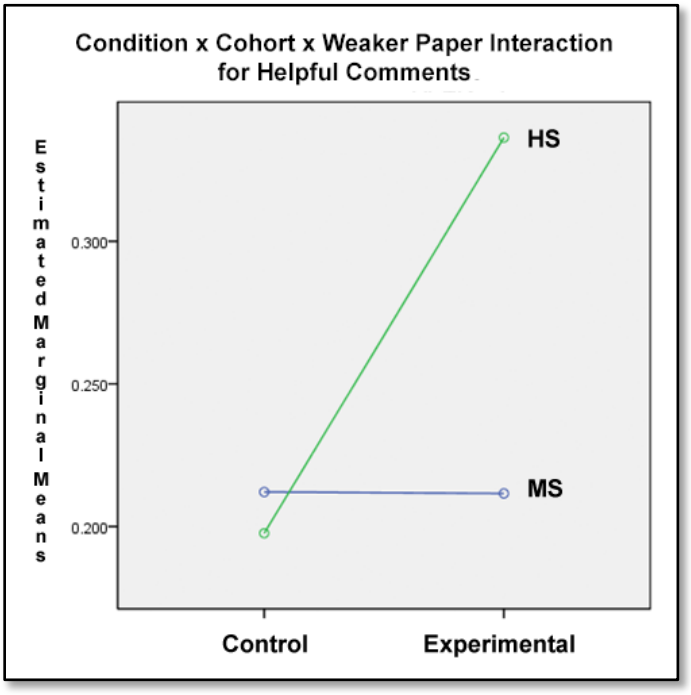


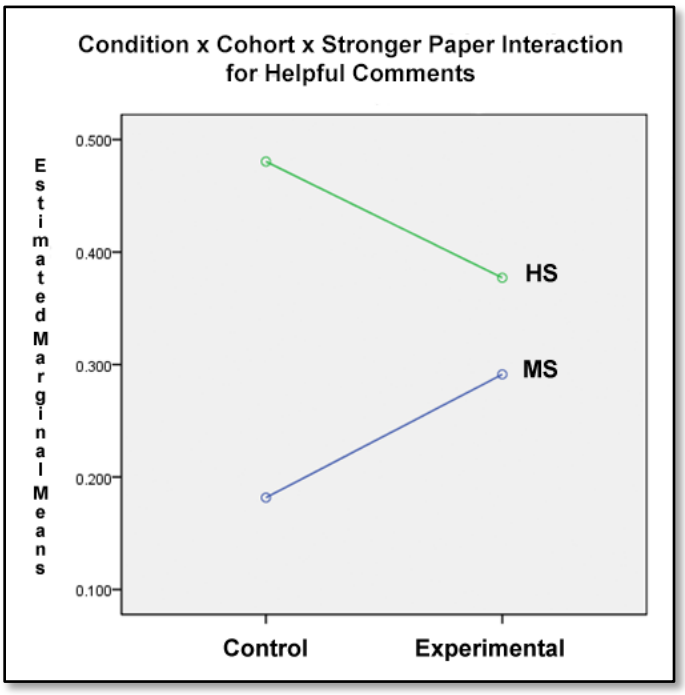
Figures 20a, b Interaction between Cohort & Paper Strength for Helpful and Unhelpful Comments





Figures21 a, b Interaction between Cohort, Experimental Condition, and Paper Strength for Helpful Comments





Discussion of the 2x2x2 Findings

Despite the fact that middle school teachers provided a somewhat greater average number of comments per essay (12.14, SD 6.401) than did their high school counterparts (9.19, SD 4.950), they offered only about half as many comments moving beyond the simple edit category (24.2%) than did those of high school teachers (41.4%). Moreover, they were remarkably less sensitive to the experimental condition of student-provided commentary added to for-submission essay. Take, for example, the feedback characteristic of focus on self-regulation. Such feedback by nature lifts the student's mind off the essay page itself and into the learning context surrounding the writing task per se, as is the case in HC6a's summative comment to Samantha Miller, "You have clearly identified what you still need to learn [about locating and documenting sources] before our research next quarter" [RCSD↑✓AH]. In the modified original analysis of FR, the omnibus experimental-versus-control effect size for self-regulation feedback was medium, at partial $\eta^2 = 0.088$. High school teachers, however, accounted for most of this effect. In fact, had the study simply focused on the feedback practices of the host district's high school teachers, the effect size would have been much greater, partial $\eta^2 = 0.112$; $F(1, 95) = 11.760$, $p = 0.001$)—not quite a large effect, but considerably greater than that for middle school teachers alone, partial $\eta^2 = 0.056$; $F(1, 112) = 6.466$, $p = 0.012$.

Similarly differentiated were middle- and high school outcomes for comparisons between the evaluated paper and the students' imaginable prior or successive work, a particularly helpful feedback characteristic for students who are not currently succeeding with respect to a skill's ideal criteria or with respect to the progress of other students. An easy-to-code self-referenced comparison appears in ME9b's summative comment to John Cauthron's weaker middle school paper: "I think you have a great start, but correcting some grammar issues and adding more

specific details will make your paper better. ☺" [WOCSE↑✓AH]. Here, the comparison refers both to the implied criteria of task-appropriate grammar and details and also to the probable outcome of an imaginably "better" subsequent draft from Cauthron.

The study illuminated a small-sized omnibus experimental-versus-control effect for such self-referenced comparisons, partial $\eta^2 = 0.042$; but as with self-regulation-focused comments this difference was driven largely by the responsiveness of high school teachers to the experimental condition. Had the study focused only on the feedback practices of high school teachers, the results would have demonstrated a medium-sized effect, with partial $\eta^2 = 0.076$; $F(1, 95) = 7.630$, $p = 0.007$. By contrast, among middle school teachers, there was no significant experiment-versus-control effect whatsoever; $F(1,112) = 0.136$, $p = 0.713$). And in fact, not only were high school teachers responsive to the experimental condition in a quantitatively measurable way, they were simply altogether much more inclined to make these sorts of comparisons regardless of experimental condition, such that the study's largest effect size proved to be the difference between middle- and high school teachers in their tendency to make these comparisons, partial $\eta^2 = 0.148$.

Generally speaking, even when they weren't responsive to the experimental conditions, high school teachers were simply more likely to engage the students' thinking with ideas that pressed far beyond the simple mechanics of "good writing" or a "correct essay" and into issues pertaining to purpose or effectiveness, or even their thought-lives beyond the paper itself. And this difference wasn't merely quantitative. For both coders during the review process, the experience of reading high school papers versus middle school ones was much like the difference between watching *The Wizard of Oz* in its black-and-white versus color scenes. Only with comparative rarity did we encounter in middle school papers a range of personal, engaging

comments that were fairly common among high school papers, as when MC24 first asked Samantha Miller, "Are you aware that a program in the library will automatically cite your sources if you just put in the info?," and then wrote in her summative notes, "You write well but seem unsure of what your LMC has to offer." These sorts of comments are highly important, for rising above a mere reactionary gestures to the paper-at-hand, they function as rather personalized invitations to subsequent learning.

Summary of the 2x2x2 Findings

In light of the preliminary findings about how middle and high school teachers scored essays differently, the study retraced its steps to illuminate meaningful differences in their feedback practices. According to the results, high school teachers—though prone to "crankiness" when confronted by comparatively poorer papers—seemed more inclined to give richly complex feedback than are middle school teachers.

- While slightly less inclined than middle school teachers to focus on the paper *per se* (partial $\eta^2 = 0.022$), high school teachers are more attuned to providing beneficial comments related to students' self-regulation (partial $\eta^2 = 0.042$).
- Similarly, while somewhat less likely than middle school teachers to make comparisons to the stipulated criteria for success (partial $\eta^2 = 0.021$), high school teachers are much more likely to imagine possible comparisons between a student's present draft and previous/successive tasks (partial $\eta^2 = 0.148$)
- Further, high school teachers exhibited a greater tendency to provide evaluative rather than descriptive comments (partial $\eta^2 = 0.048$). This was a double-edged sword, however, as not all of the evaluative comments were affirming. In practice, both data-coders found negative evaluative comments frequently to be almost

toxic in their effect on a teacher's overall communicative approach.

- Seemingly because they were less engaged with providing notations for simple edits than in other aspects of feedback, high school teachers offered a greater proportion of positive-valence comments (partial $\eta^2 = 0.061$). Perhaps partially for the same reason, however, they presented a greater proportion of negative-valence comments, as well (partial $\eta^2 = 0.014$).
- Although sometimes comparatively too prone to give tonally unhelpful comments than were middle school teachers (partial $\eta^2 = 0.012$), high school teachers were much more likely to provide tonally helpful comments as well (partial $\eta^2 = 0.062$).
- Middle- and high school teachers' clarity and specificity remained constant across cohorts.

Strength of Feedback—An Exemplar

If feedback is to be markedly beneficial to the student, it must be rich its characteristics of *focus* (on the work, on the process of completing the work, on the student's self-regulation), *comparison* (to meaningful criteria, to the student's own previous or successive attempts), *valence* (illuminating the positive and offering guidance where targets have been missed, *clarity*, *specificity*, and *tone* (inviting and inspirational rather than bossy). And while it cannot be too thinly developed in terms of the number of comments given, such feedback ought to avoid presenting the student with an overload of information at any given point in the writing/revision process. Too much of a good thing is not so good at all (Brookhart, 2008).

Within this study's data set, both coders found one participant-teacher—MC2, hereafter to be spoken of by the pseudonym *Elizabeth*—who could serve as a model for her peers in the

host district. Elizabeth's work is so strong that we have in retrospect often suggested to each other that the best professional development move we could have made for the current academic year would have been to pin photocopied sheets of her work to the bulletin boards above our desks at school as a reminder of the feedback goals to which we should aspire every time we sit down with a stack of submitted essays. Figures 21 and 22 present Elizabeth's comment sets; these figures will serve as anchor pieces for a brief discussion on optimal feedback practices as defined by Brookhart (2008) and as witnessed in the study. This discussion will be rounded out by other ideal and problematic examples from the dataset, with the purpose of making pedagogical suggestions at this study's close.

The first remarkable characteristic of Elizabeth's work is that she hasn't delivered that great a number of comments—only 6 per paper, well below the average 12.14 (SD 6.401) of her middle school peers (SD) and the 10.78 average (SD 5.962) of the entire participant pool. From this perspective, Elizabeth is in about the 6th percentile among her middle school colleagues in terms of comment frequency and the 5th percentile overall. Yet of her 6 total comments on each paper, a strikingly high proportion involve valence and tone—5 on John Cauthron's weaker text and 4 on Roger Hengst's stronger text; in other words, 83% and 67% of the comments on each paper, respectively, rise above mere "simple edits." And here again, Elizabeth is something of an outlier—in the 93rd percentile among middle school teachers and the 89th percentile overall for the weaker paper, in the 87th percentile among middle school teachers and 81st percentile overall for the stronger paper. Clearly—as both these numbers and Figures 22 and 23 themselves display—she has opted for a "less is more" strategy, trading the volume of total comments for the richness of each communicative gesture. And in crafting her low-frequency, high-impact feedback, Elizabeth offers educators much to consider in their own practices.

Figure 22 Elizabeth's Feedback to John Cauthron

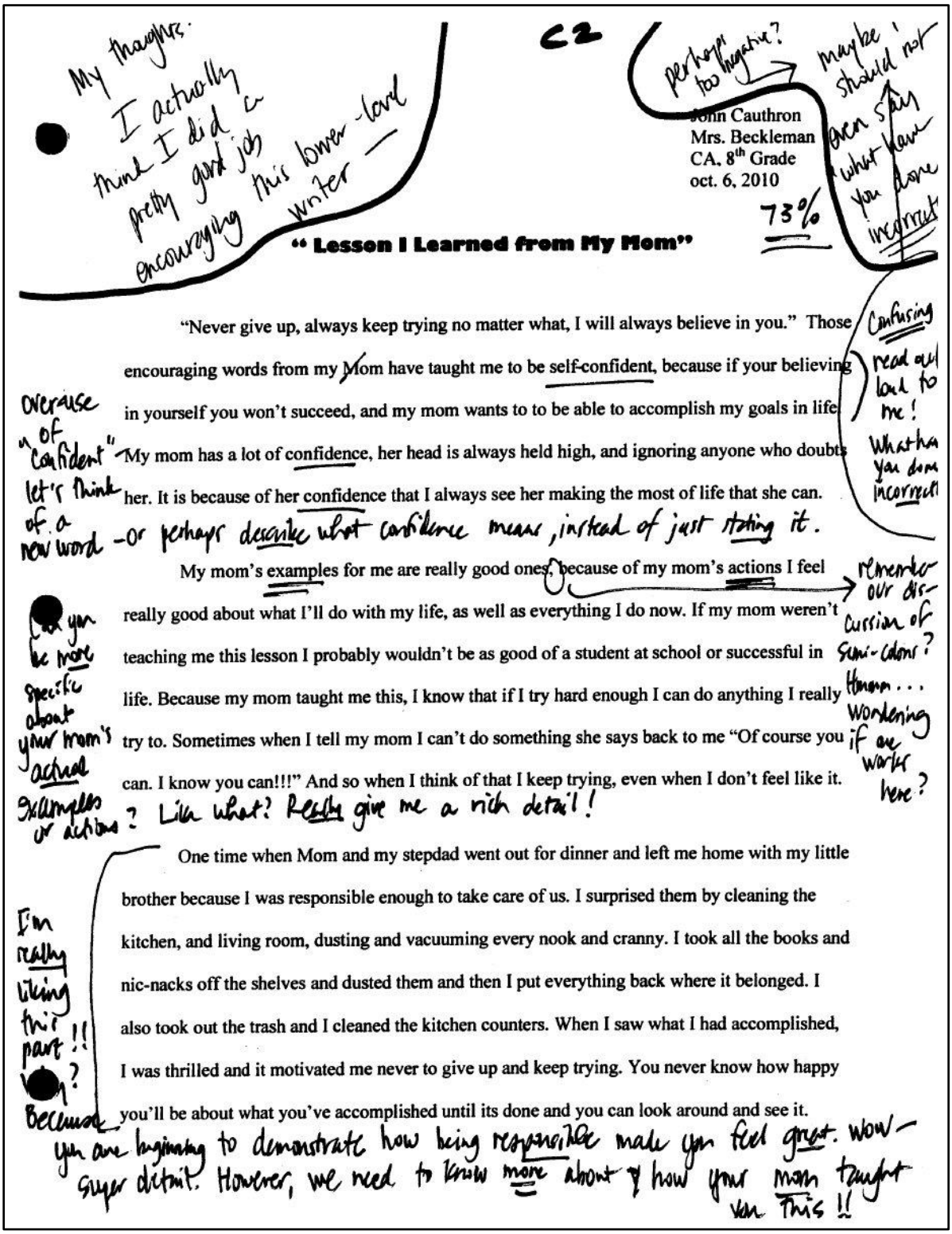
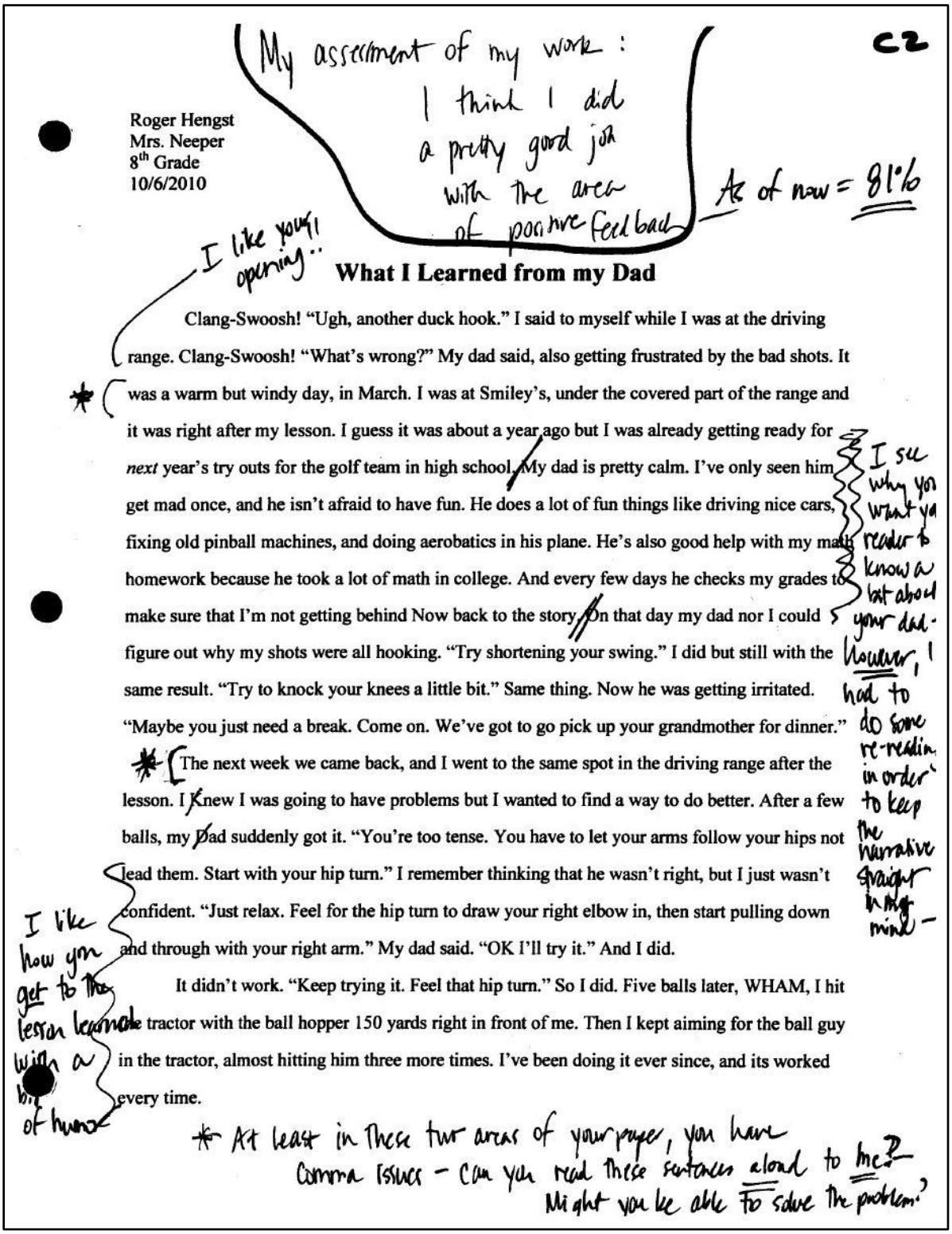


Figure 23 Elizabeth's Feedback to Roger Hengst



My assessment of my work: I think I did a pretty good job with the area of positive feedback

C2

Roger Hengst
Mrs. Neeper
8th Grade
10/6/2010

As of now = 81%

I like your opening..

What I Learned from my Dad

Clang-Swoosh! "Ugh, another duck hook." I said to myself while I was at the driving range. Clang-Swoosh! "What's wrong?" My dad said, also getting frustrated by the bad shots. It was a warm but windy day, in March. I was at Smiley's, under the covered part of the range and it was right after my lesson. I guess it was about a year ago but I was already getting ready for *next* year's try outs for the golf team in high school. My dad is pretty calm. I've only seen him get mad once, and he isn't afraid to have fun. He does a lot of fun things like driving nice cars, fixing old pinball machines, and doing aerobatics in his plane. He's also good help with my math homework because he took a lot of math in college. And every few days he checks my grades to make sure that I'm not getting behind Now back to the story. On that day my dad nor I could figure out why my shots were all hooking. "Try shortening your swing." I did but still with the same result. "Try to knock your knees a little bit." Same thing. Now he was getting irritated. "Maybe you just need a break. Come on. We've got to go pick up your grandmother for dinner."

The next week we came back, and I went to the same spot in the driving range after the lesson. I knew I was going to have problems but I wanted to find a way to do better. After a few balls, my dad suddenly got it. "You're too tense. You have to let your arms follow your hips not lead them. Start with your hip turn." I remember thinking that he wasn't right, but I just wasn't confident. "Just relax. Feel for the hip turn to draw your right elbow in, then start pulling down and through with your right arm." My dad said. "OK I'll try it." And I did.

It didn't work. "Keep trying it. Feel that hip turn." So I did. Five balls later, WHAM, I hit the tractor with the ball hopper 150 yards right in front of me. Then I kept aiming for the ball guy in the tractor, almost hitting him three more times. I've been doing it ever since, and its worked every time.

I see why you want ya reader to know a lot about your dad. However, I had to do some re-readin in order to keep the narrative straight in my mind -

I like how you get to the lesson learned with a bit of humor

* At least in these two areas of your paper, you have comma issues - Can you read these sentences aloud to me? Might you be able to solve the problem?

What does her good feedback look like?

First, although it focuses unswervingly on the works themselves—including all 6 comments on both papers—Elizabeth's feedback frequently draws the students' attention also to the processes by which they are completing their work, as well as to matters of self-regulation. Spotting Roger Hengst's lack of control with commas, Elizabeth doesn't simply correct the errors and move on. Instead she illuminates a pair of sentences in which comma errors occur and then suggests, "At least in these two areas of your paper, you have comma issues—can you read these sentences aloud to me? Might you be able to solve the problem?" [WORSCD↑✓AH]. The move is both magnanimous and academically efficient in a Deweyan sense, on one hand inviting Hengst to engage in collaborative problem-solving and implying that he may in fact be able to solve the problem himself, while on the other hand suggesting that an alternative process of reading through his own words might illuminate the path from his existing draft to one that better reflects the transfer of comma principles from academic exercise to actual writing. Similarly, where Cauthron has incorporated a comma splice into his piece, Elizabeth circles it and notes, "Remember our discussion of semi-colons? Hmm . . . wondering if one works here" [WRCD↑✓AH]. While it is possible that Cauthron will merely change the comma to a semicolon without thinking the matter through—which would much more likely be the case had Elizabeth simply inserted the semicolon herself—it is also quite possible that he will think back to the hypothetical discussion of semicolons, remember their purposes, and even remember how they can be used to help avoid comma splices.

Frankly, Elizabeth's are remarkable moves, easily spotted as "superior" ways of alerting students to problems, but challenging to perform in the day-to-day grind providing feedback. It is all too easy simply to correct comma errors and move on, particularly if we are somewhat given

to a compulsive need to mark every error we see, as seemed the case in multiple examples from the data set (e.g., MC25b, ME5b, HE1b). HC10, for example, made seven comma corrections to Paula Healey's paper as part of an almost mind-boggling 31 comments overall (or one comment every 11.6 seconds). While she might be an excellent copy editor, with 29 of her 31 notations being simple edits, HC10's high volume of marks comes at the expense of selectivity of response and warmth in tone—both of which Brookhart has suggested as characteristics of the most educationally valuable feedback. At her most generous, HC10 barely has left herself the time to write next to sentence two, "Good idea to use concrete images" [WCE↑✓AH]. And even on Samantha Miller's stronger paper, the 11 simple edits prevent HC10 from offering much by way of a positive, personal commentary. HC10 herself recognizes this weakness in her approach to Miller's paper, noting, "While I don't think my feedback has an overly negative note, there is more critical feedback than positive feedback even though this is a good piece."

Elizabeth's work is far from suffering this same imbalance between flatly editorial and roundly engaging comments. And if the first remarkable feature of its contents is its richly comprehensive focus made possible by selectivity in the comments offered, a second noteworthy aspect lies in Elizabeth's tendency to draw the student's attention to the criteria of "good writing" in ways that are affirming and empowering rather than flat, off-putting, or even denigrating. An example of this tendency lies in her response to a problem in Hengst's opening paragraph, which is filled with so many interesting ideas about his father that they take on a life of their own, creating a somewhat confusing introduction [Appendix A]. Teachers in the control condition offered a range of responses to this problem, as demonstrated in the following,

representationally fair selection of comments selected from the control group[23]:

- "Getting a little distracted" [WCD🡻✓~~AH~~] (MC5a)
- "A little disconnected, not needed or should be integrated better into the story" [WCD🡹✓AH] (MC7a).
- "Divide into 2 ¶s" [WCD✓~~A~~] (MC16a).
- "Does this fit" [WRCD🡹✓AH] (MC20a).
- "It takes me a long time to figure out what you're talking about" [WCD🡻✓~~AH~~] (MC21a).
- "Unclear organization. When talking about Dad, make smooth transitions" [WCD🡹✓AH] (MC22a).
- "Off topic & too much info for an introductory paragraph" [WCD🡻✓~~AH~~].

Several teachers in the experimental group also offered feedback to Hengst's problem with organization and focus. These responses—as this study's hypotheses implied would be likely—often provided greater vividness and warmth than did those of the control group. For whereas members of the control group were only responding to the essay itself, experimental-group participants frequently seem often to have initiated their comments as a strongly interpersonal response to Hengst's own marginal comment, "Q1: Do I get off track here, or is it good?"

---

[23] It is worth noting here that our coding of feedback does not always translate well outside of the comments' original contexts. In certain contexts, for example, ME1's comment "Seems a little random" (ME1a) might very well have received something akin to a [WCD🡻?~~AH~~]. But this particular comment, written almost in the shadow of Hengst's own question about the organization, acts in the manner of a brief but cheery bit of dialogue *with* the student *about* the paper. Context isn't everything, but it accounts for much.

[Appendix B]; in other words, they were responding to more-or-less directly to Hengst himself, not to the essay-as-text:

- "If you have to say, 'Now back to the story,' You're getting off track" [WRCD↑AH] (ME2a).
- "Why do you need this background on the dad?" [WCD↑✓AH] (ME6a).
- "The information is great. The organization is a bit confusing. Could we re-work it?" [WOCD↑✓AH] (ME8a).
- "While it gives an idea of what your dad is like, I think it gets a bit off topic" [WRCD↑✓AH] (ME9a).
- "Divide up your descriptions of the scene and setting & those of your dad for more structured organization. Jumps back & forth, hard to follow where you're headed in 1st ¶" [WCD↑✓AH] (ME13a).
- "Yes, you're off track—this info isn't necessary" [WRCD↑✓AH] (ME30a).

In the examples above, where teachers seem to have been guided by Hengst's own stated concerns about his paper, they seem to have reduced their tendency to offer "random and disparate criticisms of the formal properties of a text; in effect, notes to a paper" (Fuller, 1987, p. 308) while increasing their propensity to speak "to a person," that is, to speak with interpersonal warmth to the student-author behind the text.

Elizabeth herself was in the control group, so her response was not in any way guided by the ability to respond to Hengst's own stated concerns about his work. But despite that contextual deficit, there is much more in common between her approach and that of many experimental-group participants than with the remainder of the control group. First, she introduces the problem of disorganization by clearly positioning herself as an interested reader,

sympathetic to Hengst's desire for self-expression: "I see why you want your reader to know a bit about your dad." Only after this initial step does she add, "However, I had to do some re-reading in order to keep the narrative straight in my mind" [WCD↑✓AH]. Because of her first step, and perhaps too because she seems to locate the organizational problem "in my mind" rather than in Hengst's on thought processes, Elizabeth seems to have crafted as gently non-threatening a comparison between Hengst's current level of success and an understood criterion of good writing as might be possible.

A third success in Elizabeth's approach runs somewhat against the current of Brookhart's (2008) advice regarding evaluation. Brookhart writes "Students are less likely to pay attention to descriptive feedback if it is accompanied by judgments, such as a grade or an evaluative comment. Some students will even hear 'judgment' when you intend description" (p. 24). Elsewhere, Brookhart even goes so far as to write that "telling students the work is 'good' or 'bad'" is an example of "bad feedback function" (p. 35). Certainly, there were many examples in this study's dataset of evaluative comments to be avoided. Such comments frequently appeared as part of a summative discussion of the paper's holistic merits. Responding to Paula Healey's essay about locating useful sources, HC15 wrote, "Sentences just tacked together with no thought to organization" [WCE↓✓A~~H~~], and HC16 noted "Opening sets context but is bland/generic" [WCE↓✓~~AH~~]. While it is possible that Healey has thrown together her essay at the last moment, giving little thought to organization or the need to capture her readers' interest, such evaluative comments as these about an essay's written elements are not likely to be beneficial to the learning process.

Even worse is MC20's response to the personal anecdote John Cauthron has chosen to explain what he has learned from his mother: "Example not that good (cleaning)?"

[WCE🡻✓A~~H~~] (MC20b). Such a response from an educational professional seems unfortunate and counterproductive to the student's learning process. Similarly worrisome are the following summative comments from MC29 in response to Cauthron's work: "Content—lacking (no specific examples); we see no evidence of what mom has done" [WCE🡻✓~~AH~~], and "Word Choice—lacking" [WCE🡻✓~~AH~~]. Evaluative comments such as these may be technically accurate, but they fall short of ideal feedback both in their lack of specific guidance toward improvements and also in the tone they take with another human being's minor child. We must be careful to avoid such responses. Similarly, we must extend our carefulness even with respect to what may pass itself off to experts as "technical language" but which is likely to be heard as excessively negative by our students. How often have we written "awkward sentence" [WCE🡻?~~AH~~] or "choppy" [WCE🡻?~~AH~~] without pausing first to think how these sentences must sound to at least the more writing-averse of our students, or even how little useful information such notations provide?

When Elizabeth offers evaluation to her students, her work is cut from a different cloth altogether than that of the examples above. Her work reads as being carefully crafted so as to engage the student's interest and trust. Sometimes her evaluative comments are quite simple and somewhat general, as with the "I like your opening!!" statement attached to the first three lines of Hengst's essay. At other times, they are more specific. "I like how you get to the lesson learned," she tells Hengst at the end of his paper, "with a bit of humor" [WCE🡹✓AH]. But in every case her evaluative comments are emotionally courteous and kind. And this emotional courtesy appears even when Elizabeth is delivering comparatively bad news. In response to Cauthron's final paragraph—the same paragraph prompting MC20's "Example not that good" comment—Elizabeth has initiated her response in a characteristically upbeat manner: "I'm really

liking this part!! Why? Because you are beginning to demonstrate how being responsible made you feel great. Wow—super detail." Only after establishing this common ground with her student does Elizabeth point him in the direction of needing better writing outcomes: "However, we need to know more about how your mom taught you this!!" [WCE↑✓AH]. Similarly, when delivering what Hengst might consider to be bad news—that he has earned only an 81% rather than his predicted *A*—she does so while deftly adding a friendly invitation to further revision: "As of now."

Apart from her tendency toward evaluative comments, Elizabeth's feedback otherwise conforms to Brookhart's research-based suggestions. She quite nearly always, for example, crafts her language in such a way as to generate a positive valence and a helpful tone, as when noting Cauthron's need for greater variety in his diction: "Overuse of 'confident'—let's think of a new word—or perhaps describe what confidence means, instead of just stating it." [WOCD↑✓AH]. By offering (not merely one but) two suggestions for fixing the problem Elizabeth has avoided the trap of offering a negative critique without also proposing a path toward improvement. And by not simply fixing the problem herself, she has fostered the further development of Cauthron's own problem-solving processes. She has shown respect to Cauthron as the primary agent of his own education rather than simply editing his work for him.

This respectful, fully developed response style appears not only in areas that might be easily be addressed with valence-neutral "simple edits," but also where thornier problems have arisen in the content or organization of a piece. Cauthron's second paragraph, for example, is fairly flat and lacking the color that vivid details might provide. Where many of us might simply write, "Needs vividness," or "Add details," Elizabeth's selectivity with respect to the volume of her comments allows a more fully fledged response: "Can you be more specific about your

mom's actual examples or actions? Like what? Really give me a rich detail!" [WCD↑✓AH]. And even when she comes close to sounding frustrated with Cauthron's lack of organization in paragraph one, Elizabeth tempers her frustration with an invitation to individual assistance and a call for further reflection: "Confusing, read out loud to me! What have you done incorrectly?" [WCDOR↑✓AH].[24]

It is for the reasons described above—as much as for her consistent clarity and specificity in communication, and even for her sense that subsequent drafts might receive substantially different feedback and different grades once the authors' "big issues" have been resolved—that Elizabeth stands at the head of the class in the host district's faculty of secondary ELA teachers. And it is with her work as something of an exemplar that this study now turns toward the home stretch of discussing just exactly what the experimental condition seems to have at times to have evoked from participating teachers. For when they appear to have responded to this condition, participating teachers seem to more closely approximated Elizabeth's nearly ideal approach to feedback.

Strength of Feedback—The Effects of Student Commentaries

The two points at which this study provided significant results stemming from the major variable of interest—the simulated student comments—are worthy of brief consideration before closing this work.

---

[24] It is worth noting here that although both coders reached the agreement on the positive valence and helpfulness of this comment—the invitation to "read out loud to me"—being central to our judgment, Elizabeth herself was not self-congratulatory, reflecting that she was "perhaps too negative?" and that maybe she should abandon the language of *incorrectness* when offering feedback to students.

Focus on students' self-regulation (FR) improved by the magnitude of a medium-sized effect (partial $\eta^2 = 0.087$). Much of the change is simply inherent to the question-and-response landscape made available by the comments themselves. If a student asks a question even as simple as John Cauthron's, "Should I say what I want to do or is that off-track?" [Appendix D], a great proportion of the meaningful responses from teachers will involve elements focused not only on the essay but also on what the student does "to monitor and control their own learning" (Brookhart, 2008, p. 21). In this manner, ME1's economical "Yes, great support" [WRCE↑✓AH] speaks both to the essay's performance and also to the Cauthron's perception of what that performance may or may not have accomplished. So too does ME2's response to Roger Hengst's question, "Do I get off track here, or is it good?" [Appendix B]: "If you have to say, 'Now back to the story,' you're off track" [WRCD↑✓AH]. Similarly, where Paula Healey has written, "Q1: I don't know how to fix all my p[assive] v[oice] [Appendix H], HE11's response provides a means by which Healey can monitor verb structures in this and other papers: "Look for 'to be' verbs and infinitives like the ones labeled; put the noun before the verb" [ORCD↑✓AH].[25] The effect of student-provided annotations isn't really magic so much as a deliberate provocation of a certain variation of communicative utterance. But because this particular variation has been shown to be a rather powerful one with respect to good feedback practices, perhaps we should do whatever possible to encourage its use.

While the proportion of comments focused on self-regulation improved by a medium

[25] Admittedly, this is incomplete and perhaps misleading advice regarding the amelioration of passive voice structures, but it does nevertheless serve the purpose of illuminating the sorts of interchanges provoked by the addition of student-authored commentaries.

effect size under the experimental condition, a minor increase in participant-teachers' likelihood of drawing comparisons between the student's current work and imagined previous and/or successive attempts (S) was also propelled by student-applied comments (partial $\eta^2 = 0.042$). At times these comparisons were merely implied, but at other times they were more or less explicitly rendered. HE13, for example, implies this sort of comparison in a response to Healey. Healey has asked the question, "How do I make [the paper] jump around less?" [Appendix H], to which HE13 replies, "You are correct that your ideas jump around. Come back to a central focus and then provide details to support that" [RCSD↑✓AH].[26] The self-referential element here is implied but clear enough; if the student completes the requested task well, the next draft will be an improvement over the current one. A similar implied comparison appears in HE22's slightly better response to the same question: "Break it down step-by-step; organize w/ chart/bullets to increase flow" [WORCSD↑✓AH]. By way of contrast, a rather explicit self-referenced comment appears in HE22's comment to Paula Healey where Healey has inserted "needed" into line seven of her text [Appendix H]: "Good change" [WRCSE↑✓AH].

This last is a small move on HE22's part, provoked by a similarly small move from Healey. And yet—in a way that not even the qualitative coding itself can reveal—this small move has provided Healey with feedback not only focused on the task she has attempted, and doing so in a positive, helpful sort way, but with specific reference to a concern clearly pertaining to Healey herself. Because Healey was interested enough to add the last-minute

[26] In retrospect, the code given here should more likely have been [WORCSD↑✓AH]; clearly the focus is not simply on self-regulation but also on the work itself and a process the student might employ to improve the existing draft.

comma, HE22's comment enjoys quite a bit of pedagogical leverage. It has become an important comment precisely because it is here that Healey's attention has already been drawn to the question of an improved text. That bit of targeted relevance is important, perhaps important enough to merit a change in the pedagogical habits by which we collect and respond to so-called "final drafts" of student writings.

Pedagogical Implications

The changes in teacher feedback generated by this study's student-supplied annotations were only modest in magnitude, but it is worth noting that these effects appeared under the condition of teachers receiving and evaluating a simulated assignment written by non-existing students. Moreover, the participating teachers were presented no hints about just what the study was hoping to provoke. Given this study's results, however, it wouldn't be surprising to expected that *informed* teachers who were interested in enriching the quality of their feedback might achieving strikingly powerful changes in their feedback habits as a result of implementing strategies not unlike those implied by this study's parameters. Specifically, teachers might foster such changes by incorporating a few simple steps into the submission and assessment routines.:

- Follow Hattie and Timperley's (2007) advice in assigning writing tasks for which the "goals are specific and challenging but task complexity is low" (p. 85-86). Toward this end, it may very well be more effective for most students' growth to assign a high-frequency series of well-focused, one-page tasks with rich self- and teacher-authored feedback than two or three major papers per semester without much else in between.
- Recognizing that students' self-evaluations are "critically important" to their growth as thinkers and writers (Schunk, 1990, p.164), and that they can be

accomplished effectively as early as elementary school (Andrade, Du, & Wang, 2008), carve out time for students to evaluate, annotate, and even correct their own texts before each first submission. Provide them with evaluative rubrics—the same ones used by the teacher—to aid them in accomplishing this work. Then, give students credit for spotting their own errors, even on a last-minute basis. Further, ask students to predict the grades they have earned and to justify why their grades should be neither higher nor lower than these predictions; teachers need not make any undue promises to honor these predictions, but they can gain marvelous insights by the information students provide.

- On peer review and graded submission days, hold students responsible for providing self- and peer-feedback as robust as our own. Provided with Brookhart's paradigm and meaningful models of good feedback, they may be capable of much more than we initially suppose.
- Rather than starting from preconceived—or even curriculum-stipulated—ideas about what students out to be learning in this unit or this task, ask students to jumpstart the teacher's feedback process by writing one or two specific comments or questions about the suspected problems in their own texts for which they are most interested in learning solutions.
- Respond to students' papers *and* comments generously; where the teacher and students' evaluations of the paper's merits differs significantly, take the opportunity to engage the student in follow-up conversations to learn better the nature of the disconnect. Over time, look for that disconnect to diminish.
- Give students the opportunity to revise their teacher-evaluated work for improved

> outcomes. As Brookhart writes, "It is not fair to students to present them with feedback and no opportunities to use it. It is not fair to students to present them with what seems like constructive criticism and then use it against them in a grade or final evaluation" (p. 2). From even a common-sense point of view, it's reasonable to imagine that when students enjoy no meaningful opportunity to improve their grade-wise outcomes on the heels of our feedback, they may well not see the value in returning to an already graded work. On the other hand, opportunities for revision may be supportive not only of students' sense of fairness or their willingness to reconsider a graded draft. For when our richly supportive feedback travels hand-in-hand the opportunities for improved credit through revision, we may be affording ourselves a stronger incentive to give papers accurate rather than inflated grades—when the "grade" is not final, we don't have to hedge our bets.

The time commitment for such a routine seems extravagant at first, but it may actually prove to be educationally efficient in the Deweyan sense (Dewey, 1915), where efficiency is not judged in columns and rows, or in seat time, but in the uptake of several discrete strands of learning in order to perceive and resolve better a complex problem—in this case, the problem of writing well.

In my own experience as a classroom practitioner, working with essays in the manner described above has proven invaluable to my sense of professional satisfaction, reducing my stress over assigning grades and increasing my sense of using feedback to generate "conversations" with my students about their learning, rather than simply providing somewhat fragmentary editorial comments.

Recommendations for Further Study

The partial support of this study's central hypotheses may to warrant further investigation into the alteration of assessment practices toward a model implied by this study's simulated student comments. It would be interesting to learn if this study's findings hold true across various text types or a wider range of exemplars that more rigorously tap into the sometimes bewildering array of success and failure that comprises every stack of submitted essays. Perhaps, too, further investigation might consider replicating this study's feedback approach in other contexts, particularly those involving authentic, classroom-based settings.

Along this latter line, perhaps the most valuable follow-up to this work might involve a cohort of volunteers, trained in best practices for feedback and then tracked during a year-long mixed-methods study. One group would simply use Brookhart's work as a guide for their practices, while the other would use Brookhart, supplemented with the practice of actually having students submit self-reflective commentaries on their texts as in the manner of the current study. Such an approach could be used to investigate several questions simultaneously: (a) the quality and characteristics of teachers' feedback under the two conditions, (b) the improvement of student outcomes in writing, (c) students' sense of satisfaction with the year-long learning process, (d) teachers' sense of satisfaction with the grading and feedback process.

Conclusion

The aim of this study was to investigate whether the addition of a student's self-evaluative comments might have any of three effects on teachers' grading and feedback practices. Largely, the results were encouraging. In the presence of such comments, teachers were modestly more likely to engage in best-practices feedback, comparably likely to reach reliable conclusions about an essay's merits, and rather unlikely to succumb to grade inflation.

But as surely must be the case with most dissertations, this paper is not really much more than a conversation-propeller. For the last several years I've been engaged in discussions with my colleagues about ways to do our work better while not consigning ourselves to Saturdays and Sundays on campus, sequestered away from our families and friends. It is possible that the practice of incorporating students' self-evaluative comments into our own grading process—a practice seeming to leverage benefits anticipated by Heider's (1958) psychology of interpersonal relations—might be one such way of reclaiming our too heavily sacrificed time.

References

Allen, M. (1995). Valuing differences: Portnet's first year. *Assessing Writing, 2*, 67-90.

Allington, R. (2001). What really matters for struggling readers: Designing research-based interventions. New York: Longman, 2001.

American College of Emergency Physicians. (2011). ACEP 2011 national emergency physicians survey results. *American College of Emergency Physicians*. Retrieved May 26, 2011, from http://www.acep.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=78645&libID=78673

Ammons, R. B. (1956). Effects of knowledge on performance: A survey and tentative theoretical formulation. *Journal of General Psychology, 54*, 279-299.

Andrade, H. L., Du, Y., and Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice, 27*(2), 3-13.

Apple, M. W. (1979, 2004). *Ideology and curriculum*. 3rd ed. New York: RoutledgeFalmer.

Applebee, A. N., and Langer, J. A. (2006). *The state of writing instruction in America's schools: What existing data tell us*. Albany, NY: Center on English Learning & Achievement.

Ashford, S. J., and Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. *Organizational Behavior and Human Performance, 32*, 370-398.

Atwell, N. (1987, 1998). In the middle: New understandings about writing, reading, and learning. 2nd ed. Portsmouth, NH: Boynton/Cook.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., and Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61,* 213-238.

Bardine, B. A., Bardine, M. S., and Deegan, E. F. Beyond the red pen: Clarifying our role in the response process. *English Journal, 90*(1), 94-101.

Barlow, L., Liparulo, S. P., and Reynolds, D. W. (2007). Keeping assessment local: The case for accountability through formative assessment. *Assessing Writing, 12,* 44-59.

Bowman, M., Mahon, W., and Pogell, S. (2004). Assessment as opportunity: A conversation with Brian Huot. *Issues in Writing, 14*(2), 94-115.

Brannon, L., and Knoblauch, C. H. (1982). On students' rights to their own texts: A model of teacher response. *College Composition and Communication, 33*(2), 157-166.

Breland, H., Kubota, M., Nickerson, K., Trapani, C., and Walker, M. (2004). *New SAT writing prompt study: Analysis of group impact and reliability* (Research Report No. 2004-1). New York: College Entrance Examination Board.

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education, 8*(2), 153-169.

Brookhart, S. M. (2008). *How to give effective feedback to your students*. Alexandria, VA: ASCD.

Brookhart, S. M., & Devoge, J. G. (1999). Testing a theory about the role of classroom assessment in student motivation and achievement. *Applied Measurement in Education, 10*, 161-180.

Brookhart, S. M., & Freeman, D. J. (1992). Characteristics of entering teacher candidates. *Review of Educational Research, 62*(1), 37-60.

Burkland, J., and Grimm, N. (1986). Motivating through responding. *Journal of Teaching*

Writing, 5, 237-247.

Butler, R., and Nisan, M. (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology, 78*(3), 210-216.

Butler, D. L., and Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245-281.

CCCC Committee on Assessment. (1995). Writing assessment: A position statement. *College Composition and Communication, 46*(3), 1995, 430-437.

Chase, C. I. (1968). The impact of some obvious variables on essay scores. *Journal of Educational Measurement, 5*(4), 315-318.

Chase, C. I. (1983). Essay test scores and reading difficulty. *Journal of Educational Measurement, 20*(3), 293-297.

Chase, C. I. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement, 23*(1), 33-41.

Cherry, R. D., and Meyer, P. R. (1993). Reliability issues in holistic assessment. In B. Huot and P. O'Neill (Eds.), 2009, *Assessing writing: A crucial sourcebook* (pp. 29-56). Boston: Bedford/St. Martin's. (Reprinted from *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*, pp. 109-141, by M. M. Williamson and B. A. Huot, Eds., 1993, Cresskill, NJ: Hampton Press).

Cochran-Smith, M., and Lytle, S. L. (2006). Troubling images of teaching in No Child Left Behind. *Harvard Educational Review, 76*(4), 668-697.

College Board. (2007). Exam scoring: What an AP grade means. Retrieved November 4, 2007, from http://apcentral.collegeboard.com/apc/public/exam/about_exams/1994.html

College Board. (2010). AP English literature and composition 2010 scoring guidelines. Retrieved May 25, 2011, from http://apcentral.collegeboard.com/apc/public/repository/ap10_english_literature_scoring_guidelines.pdf

Condon, W. (2009). Looking beyond judging and ranking: Writing assessment as a generative practice. *Assessing Writing, 14*, 141-156.

Conley, D. T. (2007). Toward a more comprehensive conception of college readiness: Prepared for the Bill & Melinda Gates Foundation. Eugene, OR: Educational Policy Improvement Center.

Connors, R. J. (1981a). Current-traditional rhetoric: Thirty years of *Writing with a Purpose*. *Rhetoric Society Quarterly, 11*(4), 208-221.

Connors, R. J. (1981b). The rise and fall of the modes of discourse. *College Composition and Communication, 32*(4), 444-455.

Connors, R. J. (1986). Textbooks and the evolution of the discipline. *College Composition and Communication, 37*(2), 178-194.

Cook, T. D., and Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand McNally College Publishing.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper and L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-31). Urbana, IL: National Council of Teachers of English.

Counts, G. S. (1932). *Dare the school build a new social order*? Carbondale, IL: Southern Illinois University Press.

Crawford, M. (2008). Think inside the clock. *Phi Delta Kappan, 90*(4), 251-255.

Daiker, D. A. (1989). Learning to Praise. In C. M. Anson (Ed.), *Writing and response: Theory,*

practice, and research (pp. 103-113). Urbana, IL: National Council of Teachers of English.

Darling-Hammond, L. (2006). No Child Left Behind and High School Reform. *Harvard Educational Review, 76*(4), 642-667.

Dearborn, W. F. (n.d.). School and university grades. *Bulletin of the University of Wisconsin,* 368.

Dewey, J. (1910, 1991). *How we think.* Amherst, NY: Prometheus.

Dewey, J. (1915, 2001). *The school and society*. Mineola, NY: Dover.

Dewey, J. (1938, 1997). *Experience and education*. New York: Touchstone. Graham, S., and Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools*. A report to the Carnegie Corporation of New York. Washington: Alliance of Excellent Education. Hattie, J., & Timperley. H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Diederich, P. B. (1964). Problems and possibilities of research in the teaching of written composition. In *Research design and the teaching of English: Proceedings of the San Francisco Conference* (pp. 52-73). Champaign, IL: National Council of Teachers of English.

Diederich, P. B. (1974). *Measuring growth in English*. Urbana, IL: National Council of Teachers of English.

Diederich, P. B., French, J. W., and Carlton, S. T. (1961). *Factors in judgments of writing ability* (Research Bulletin No. 61-15). Princeton, NJ: Educational Testing Service.

Dohrer, G. (1991). Do teachers' comments on students' papers help? *College Teaching, 39*(2), 48-54.

Dragga, S. (1988). The effects of praiseworthy grading on students and teachers. *Journal of Teaching Writing, 7*(1), 41-50.

DuFour, R., and Eaker, R. (1998). *Professional learning communities at work: Best practices for enhancing student achievement*. Alexandria, VA: Association for Supervision and Curriculum Development.

DuFour, R., DuFour, R., Eaker, R., and Karhanek, G. (2004). *Whatever it takes: How professional learning communities respond when kids don't learn*. Bloomington, IN: National Educational Service.

Education Northwest. (2011). *6+1 Trait writing*. Retrieved June 2, 2011, from http://education northwest.org/traits

Elbow, P. (1973). *Writing without teachers*. New York: Oxford University Press.

Elbow. P. (1997). Grading student writing: Making it simpler, fairer, clearer. *New Directions for Teaching and Learning, 69*, 127-140.

Emig, J. (1971). *The composing process of twelfth graders*. NCTE Research Report No. 13. Urbana, IL: National Council of Teachers of English.

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71*(3), 328-338.

Freedman, S. W. (1987). *Response to student writing*. Urbana, IL: National Council of Teachers of Writing.

Freire, Paulo. (1970, 2000). *Pedagogy of the oppressed*. Trans. Myra Bergman Ramos. 30th anniv. ed. New York: Continuum.

Fuller, D. C. (1987). Teacher commentary that communicates: Practicing what we preach in the writing class. *Journal of Teaching Writing, 6*, 307-317.

Gee, T. C. (1972). Students' responses to teacher comments. *Research in the Teaching of English, 6*(2), 212-221.

Godshalk, F. I., Swineford, F., and Coffman, W. E. *The measurement of writing ability*. ETS Research Monograph No. 6. Princeton, NJ: Educational Testing Service, 1966.

Gore, L., and Lloyd, J. (2011). Want to cut costs in the ER? Pass medical liability reform: Poll of emergency physicians shows more than half order tests as protection against being sued. *American College of Emergency Physicians*. Retrieved May 26, 2011, from http://www.acep.org/Content.aspx?id=79958

Graves, D. (1975). An examination of the writing process of seven year old children. *Research in the Teaching of English, 9*, 138-43.

Grobe, W. J., and McCall, D. (2004). NCLB: Failed schools—or failed law? *Educational Horizons, 82*(2), 131-142.

Grolnick, W. S., and Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology, 52*(5), 890-898.

Harris, M. J., and Rosenthal, R. (1985). Mediation of interpersonal expectancies effects: 31 meta-analyses. *Psychological Bulletin, 97*, 363-386.

Harter, S. (1978). Pleasure derived from challenge and the effects of receiving grades on children's difficulty level choices. *Child Development, 49*(3), 788-799.

Hattie, J. A. (1999). *Influences on student learning* (Inaugural professorial address, university of Auckland, New Zealand). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.8465&rep=rep1&type=pdf

Hattie, J. A., Biggs, J., & Purdie, N. (1996). Effects of learning skills intervention on student

learning: A meta-analysis. *Review of Research in Education, 66*, 99-136.

Hattie, J. A., & Timperley. H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Heider, F. (1958). *The psychology of interpersonal relations*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hillocks, G. (1986). *Research on written composition: New directions for teaching*. Urbana, IL: National Conference on Research in English.

Hillocks, G. (2006). Middle and high school composition. In P. Smagorinsky (Ed.), *Research in composition: Multiple perspectives on two decades of change* (pp. 48-77). New York: Teachers College Press.

Hopkins, E. M. (1912). Can good composition teaching be done under present conditions? *English Journal, 1,* 1-8.

Houston, P. D. (2007). The seven deadly sins of No Child Left Behind. *Phi Delta Kappan, 88*(10), 744-748.

Horvath, B. K. (1984). The components of written response: A practical synthesis of current views. *Rhetoric Review, 2*(2), 136-156.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*(4), 549-566.

Jacoby, H. (1910). Note on the marking system in the astronomical course at Columbia College, 1909-10. *Science, 31*(804), 819-820.

Kennedy, M. M. (1998). *Learning to teach writing: Does teacher education make a difference?* New York: Teachers College Press.

Kitchen, E., King, S. H., Robison, D. F., Sudweeks, R. R., Bradshaw, W. S., and Bell, J. D.

(2006). Rethinking exams and letter grades: How much can teachers delegate to students? *CBE—Life Sciences Education, 5*, 270-280.

Klein, Alyson. (2010). Sec. Duncan: Districts need to rethink class size, salary structure. *Education Week Politics K-12 Blog*. Retrieved May 22, 2011, from http://blogs.edweek.org/edweek/campaign-k-12/2010/11/sec_duncan_districts_need_to_r.html

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284.

Kohn, A. (2011). Grading: The issue is not how but why. *Educational Leadership, 52*(2), 38-41.

KSDE. (2008). *Writing assessment information*. Retrieved from http://www.ksde.org/Default.aspx?tabid=165

Land, R. E., and Evans, S. (1987). Classroom inquiry: What our students taught us about paper marking. *English Journal, 76*(2), 113-116.

Lloyd-Jones, R., (1977). Primary-trait scoring. In C. R. Cooper and L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33-66). Urbana, IL: National Council of Teachers of English.

Locke, E. A., and Latham, G. P. (1984). *Goal setting: A motivational technique that works*. Englewood Cliffs, NJ: Prentice Hall.

Lynch, C., and Klemans, P. (1978). Evaluating our evaluations. *College English, 40*(2), 166-170, 175-180.

Marshall, J. C. (1967). Composition errors and essay examination grades re-examined. *American Educational Research Journal, 4*(4), 375-385.

Marshall, J. C., and Powers, J. M. (1969). Writing neatness, composition errors, and essay grades. *Journal of Educational Measurement, 6*(2), 97-101.

Marzano, R. J. (2000). *Transforming classroom grading*. Alexandria, VA: ASCD.

McCarthey, S. J. (2008). The impact of No Child Left Behind on teachers' writing instruction. *Written Communication, 25*(4), 462-505.

Modern Language Association. (2009). *MLA handbook for writers of research papers* (7th ed.). New York: Modern Language Association.

Murray, D. M. (1982). Learning by teaching: Selected articles on writing and teaching. Portsmouth: Boynton/Cook, 1982.

National Commission on Writing. (2004). *Writing: A ticket to work… or a ticket out: A survey of business leaders*. Retrieved July 18, 2010, from http://www.host-collegeboard.com/advocacy/writing/publications.html

National Commission on Writing. (2005). *Writing: A powerful message from state government*. Retrieved July 18, 2010, from http://www.host-collegeboard.com/advocacy/writing/publications.html

National Writing Project, and Nagin, C. (2006). *Because writing matters*. Revised and updated. San Francisco: Jossey-Bass.

Neil, M. (2003). Leaving children behind: How No Child Left Behind will fail our children. *Phi Delta Kappan, 85*(3), 225-228.

Nystrand, M. (1997). Opening dialogue: Understanding the dynamics of language and learning in the English classroom. New York: Teachers College Press.

Nystrand, M. (2006). The social and historical context for writing research. *Handbook of writing research*. Ed. C. A. MacArthur, S. Graham, J. Fitzgerald. New York: Guilford. 11-27.

O'Connor, K. (2007). *A repair kit for grading: 15 fixes for broken grades*. Boston: Pearson.

Pedersen, S., and Williams, D. (2004). A comparison of assessment practices and their effects on learning and motivation in a student-centered learning environment. *Journal of Educational Multimedia and Hypermedia, 13*(3), 283-306.

Pritchard, R. D., Jones, S. D., Roth, P. L., Stuebing, K. K., and Ekeberg, S. E. (1988). Effects of group feedback, goal setting, and incentives on organizational productivity. *Journal of Applied Psychology, 73*, 337-358.

Purcell-Gates, V. (2007). Comprehending complexity. In V. Purcell-Gates (ed.).*Cultural practices of literacy: Case studies of language, literacy, social practice, and power*. Mahwah, NJ: Lawrence Erlbaum. 197-216.

Schunk, D. H. (2003). Self-efficacy for reading and writing: Influence of modeling, goal setting, and self-evaluation. *Reading & Writing Quarterly, 19*, 159-172.

Shaughnessy, M. P. (1977). *Errors and expectations: A guide for the teacher of basic writing*. New York: Oxford University Press.

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Smagorinsky, P., and Whiting, M. E. (1995). *How English teachers get taught: Methods of teaching the methods class*. Urbana, IL: National Council of Teachers of English.

Starch, D. and E. C. Elliott (1912). Reliability of the grading of high-school work in English. *The School Review, 20*(7), 442-457.

Starch, D., and Elliott, E. C. (1913). Reliability of grading work in mathematics. *The School Review, 21*(4), 254-259.

Straub, R. (1996). Teacher response as conversation: More than casual talk, an exploration.

Rhetoric Review, 14(2), 374-399.

Straub, R. (1997). Students' reactions to teacher comments: An exploratory study. *Research in the Teaching of English, 31*(1), 91-119.

Supovitz, J. A., and Brennan, R. T. (1997). Mirror, mirror on the wall, which is the fairest test of all? An examination of the equitability of portfolio assessment relative to standardized tests. *Harvard Educational Review, 67*(3), 472-506.

Sweller, J. (1990). Cognitive processes and instruction procedures. *Australian Journal of Education, 34*(2), 125-130.

Tieje, R. E., Sutcliffe, E. G., Hillebrand, H. N., and Buchen, W. (1915). Systematizing grading in freshman composition at the large university. *English Journal, 4*(9), 586-597.

United States Department of Education. (2002). *No Child Left Behind Act of 2001*. Retrieved May 21, 2011, from http://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf

United States Department of Education. (2004). *Individuals with Disabilities Education Improvement Act of 2004.* Retrieved May 21, 2011, from http://idea.ed.gov/download/statute.html

White, E. M. (1984). *Teaching and assessing writing*. San Francisco: Jossey-Bass.

White, E. M. (1995). An apologia for the timed impromptu essay test. *College Composition and Communication, 46*(1), 30-45.

Wiggins, G., and McTighe, J. E. (2005). *Understanding by design* (expanded 2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Zeichner, K., and Wray, S. (2001). The teaching portfolio in US teacher education programs: What we know and what we need to know. *Teaching and Teacher Education, 17*, 613-621.

Appendix A: Text MSA_{clean}

Roger Hengst
Mrs. Neeper
8th Grade
10/6/2010

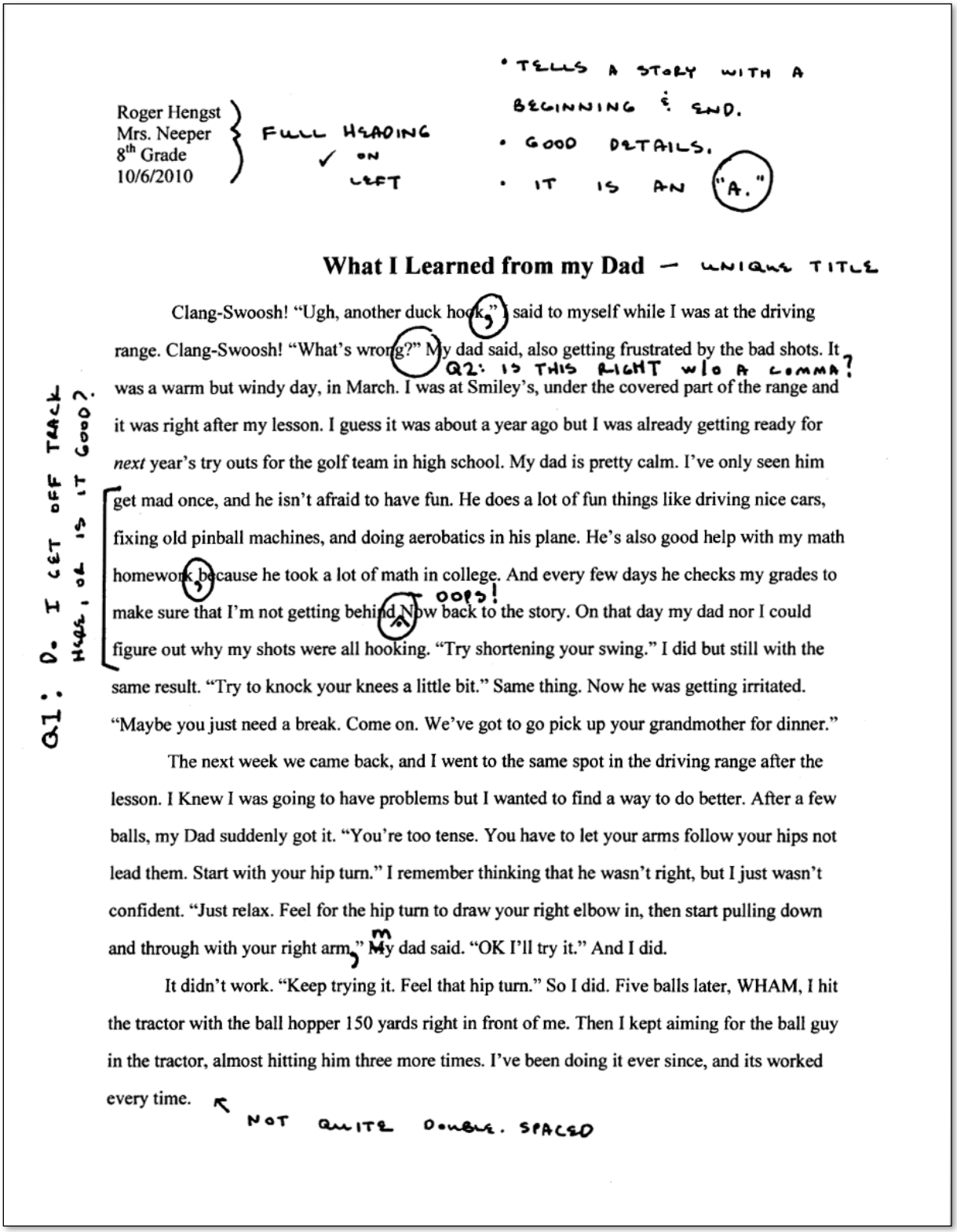
What I Learned from my Dad

Clang-Swoosh! "Ugh, another duck hook." I said to myself while I was at the driving range. Clang-Swoosh! "What's wrong?" My dad said, also getting frustrated by the bad shots. It was a warm but windy day, in March. I was at Smiley's, under the covered part of the range and it was right after my lesson. I guess it was about a year ago but I was already getting ready for *next* year's try outs for the golf team in high school. My dad is pretty calm. I've only seen him get mad once, and he isn't afraid to have fun. He does a lot of fun things like driving nice cars, fixing old pinball machines, and doing aerobatics in his plane. He's also good help with my math homework because he took a lot of math in college. And every few days he checks my grades to make sure that I'm not getting behind Now back to the story. On that day my dad nor I could figure out why my shots were all hooking. "Try shortening your swing." I did but still with the same result. "Try to knock your knees a little bit." Same thing. Now he was getting irritated. "Maybe you just need a break. Come on. We've got to go pick up your grandmother for dinner."

The next week we came back, and I went to the same spot in the driving range after the lesson. I Knew I was going to have problems but I wanted to find a way to do better. After a few balls, my Dad suddenly got it. "You're too tense. You have to let your arms follow your hips not lead them. Start with your hip turn." I remember thinking that he wasn't right, but I just wasn't confident. "Just relax. Feel for the hip turn to draw your right elbow in, then start pulling down and through with your right arm." My dad said. "OK I'll try it." And I did.

It didn't work. "Keep trying it. Feel that hip turn." So I did. Five balls later, WHAM, I hit the tractor with the ball hopper 150 yards right in front of me. Then I kept aiming for the ball guy in the tractor, almost hitting him three more times. I've been doing it ever since, and its worked every time.

Appendix B: Text *MSA*$_{annotated}$



Roger Hengst
Mrs. Neeper
8th Grade
10/6/2010

FULL HEADING ✓ ON LEFT

• TELLS A STORY WITH A BEGINNING & END.
• GOOD DETAILS.
• IT IS AN "A."

What I Learned from my Dad — UNIQUE TITLE

Clang-Swoosh! "Ugh, another duck hook," I said to myself while I was at the driving range. Clang-Swoosh! "What's wrong?" My dad said, also getting frustrated by the bad shots. It was a warm but windy day, in March. I was at Smiley's, under the covered part of the range and it was right after my lesson. I guess it was about a year ago but I was already getting ready for *next* year's try outs for the golf team in high school. My dad is pretty calm. I've only seen him get mad once, and he isn't afraid to have fun. He does a lot of fun things like driving nice cars, fixing old pinball machines, and doing aerobatics in his plane. He's also good help with my math homework because he took a lot of math in college. And every few days he checks my grades to make sure that I'm not getting behind. Now back to the story. On that day my dad nor I could figure out why my shots were all hooking. "Try shortening your swing." I did but still with the same result. "Try to knock your knees a little bit." Same thing. Now he was getting irritated. "Maybe you just need a break. Come on. We've got to go pick up your grandmother for dinner."

Q2: IS THIS RIGHT W/O A COMMA?

Q1: D. I GET OFF TRACK HERE, BUT IS IT GOOD?

OOPS!

The next week we came back, and I went to the same spot in the driving range after the lesson. I Knew I was going to have problems but I wanted to find a way to do better. After a few balls, my Dad suddenly got it. "You're too tense. You have to let your arms follow your hips not lead them. Start with your hip turn." I remember thinking that he wasn't right, but I just wasn't confident. "Just relax. Feel for the hip turn to draw your right elbow in, then start pulling down and through with your right arm," My dad said. "OK I'll try it." And I did.

It didn't work. "Keep trying it. Feel that hip turn." So I did. Five balls later, WHAM, I hit the tractor with the ball hopper 150 yards right in front of me. Then I kept aiming for the ball guy in the tractor, almost hitting him three more times. I've been doing it ever since, and its worked every time.

NOT QUITE DOUBLE. SPACED

Appendix C: Text MSB_{clean}

John Cauthron
Mrs. Beckleman
CA, 8th Grade
oct. 6, 2010

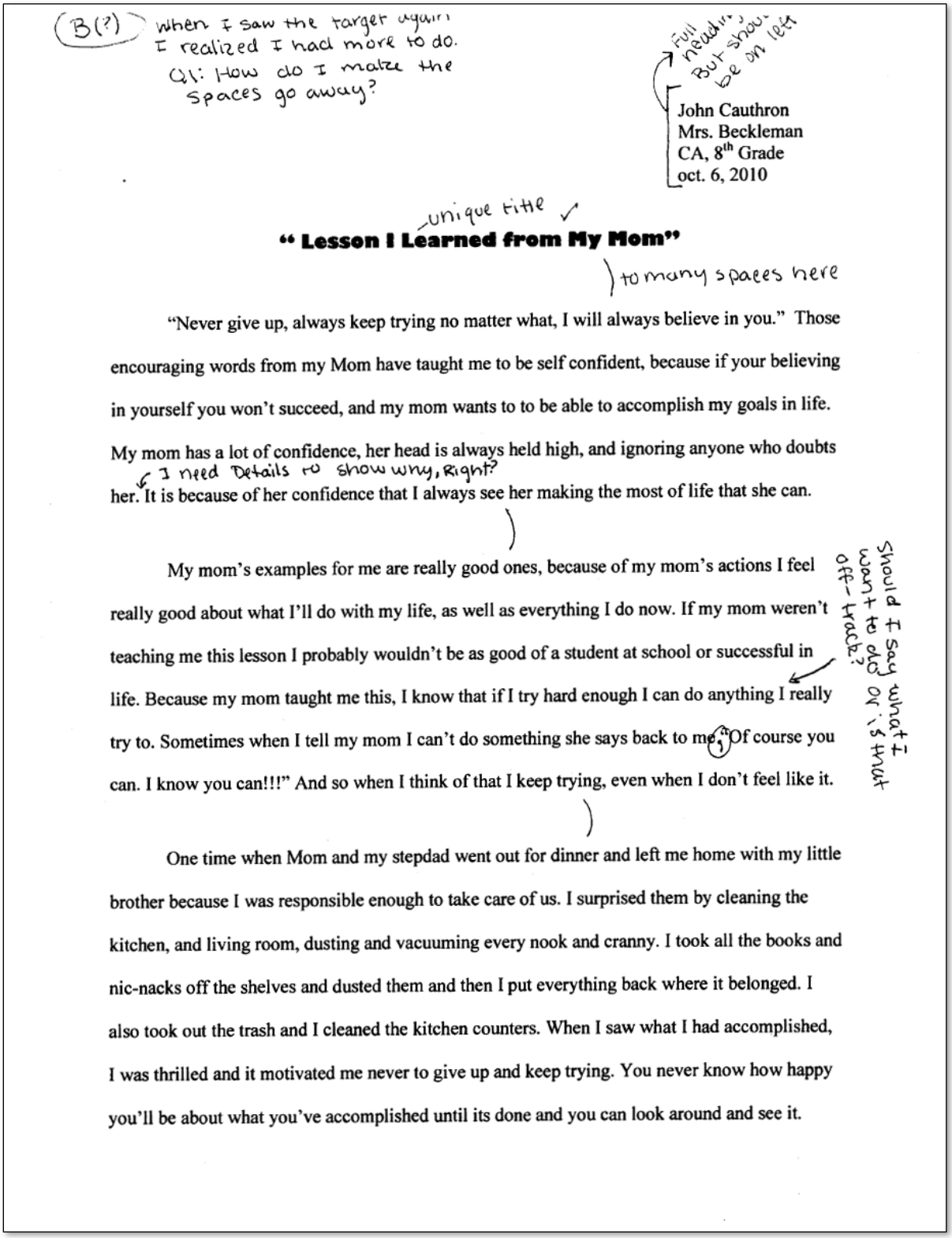
" Lesson I Learned from My Mom"

"Never give up, always keep trying no matter what, I will always believe in you." Those encouraging words from my Mom have taught me to be self confident, because if your believing in yourself you won't succeed, and my mom wants to to be able to accomplish my goals in life. My mom has a lot of confidence, her head is always held high, and ignoring anyone who doubts her. It is because of her confidence that I always see her making the most of life that she can.

My mom's examples for me are really good ones, because of my mom's actions I feel really good about what I'll do with my life, as well as everything I do now. If my mom weren't teaching me this lesson I probably wouldn't be as good of a student at school or successful in life. Because my mom taught me this, I know that if I try hard enough I can do anything I really try to. Sometimes when I tell my mom I can't do something she says back to me "Of course you can. I know you can!!!" And so when I think of that I keep trying, even when I don't feel like it.

One time when Mom and my stepdad went out for dinner and left me home with my little brother because I was responsible enough to take care of us. I surprised them by cleaning the kitchen, and living room, dusting and vacuuming every nook and cranny. I took all the books and nic-nacks off the shelves and dusted them and then I put everything back where it belonged. I also took out the trash and I cleaned the kitchen counters. When I saw what I had accomplished, I was thrilled and it motivated me never to give up and keep trying. You never know how happy you'll be about what you've accomplished until its done and you can look around and see it.

Appendix D: Text *MSB*$_{annotated}$



B(?) When I saw the target again I realized I had more to do. Q1: How do I make the spaces go away?

full heading but should be on left

John Cauthron
Mrs. Beckleman
CA, 8th Grade
oct. 6, 2010

unique title ✓

" Lesson I Learned from My Mom"

to many spaces here

"Never give up, always keep trying no matter what, I will always believe in you." Those encouraging words from my Mom have taught me to be self confident, because if your believing in yourself you won't succeed, and my mom wants to to be able to accomplish my goals in life. My mom has a lot of confidence, her head is always held high, and ignoring anyone who doubts her. It is because of her confidence that I always see her making the most of life that she can.

I need Details to show why, Right?

My mom's examples for me are really good ones, because of my mom's actions I feel really good about what I'll do with my life, as well as everything I do now. If my mom weren't teaching me this lesson I probably wouldn't be as good of a student at school or successful in life. Because my mom taught me this, I know that if I try hard enough I can do anything I really try to. Sometimes when I tell my mom I can't do something she says back to me, "Of course you can. I know you can!!!" And so when I think of that I keep trying, even when I don't feel like it.

Should I say what I want to do or is that off-track?

One time when Mom and my stepdad went out for dinner and left me home with my little brother because I was responsible enough to take care of us. I surprised them by cleaning the kitchen, and living room, dusting and vacuuming every nook and cranny. I took all the books and nic-nacks off the shelves and dusted them and then I put everything back where it belonged. I also took out the trash and I cleaned the kitchen counters. When I saw what I had accomplished, I was thrilled and it motivated me never to give up and keep trying. You never know how happy you'll be about what you've accomplished until its done and you can look around and see it.

Appendix E: Text HSA_{clean}

Samantha Miller

Mr. Stapleton

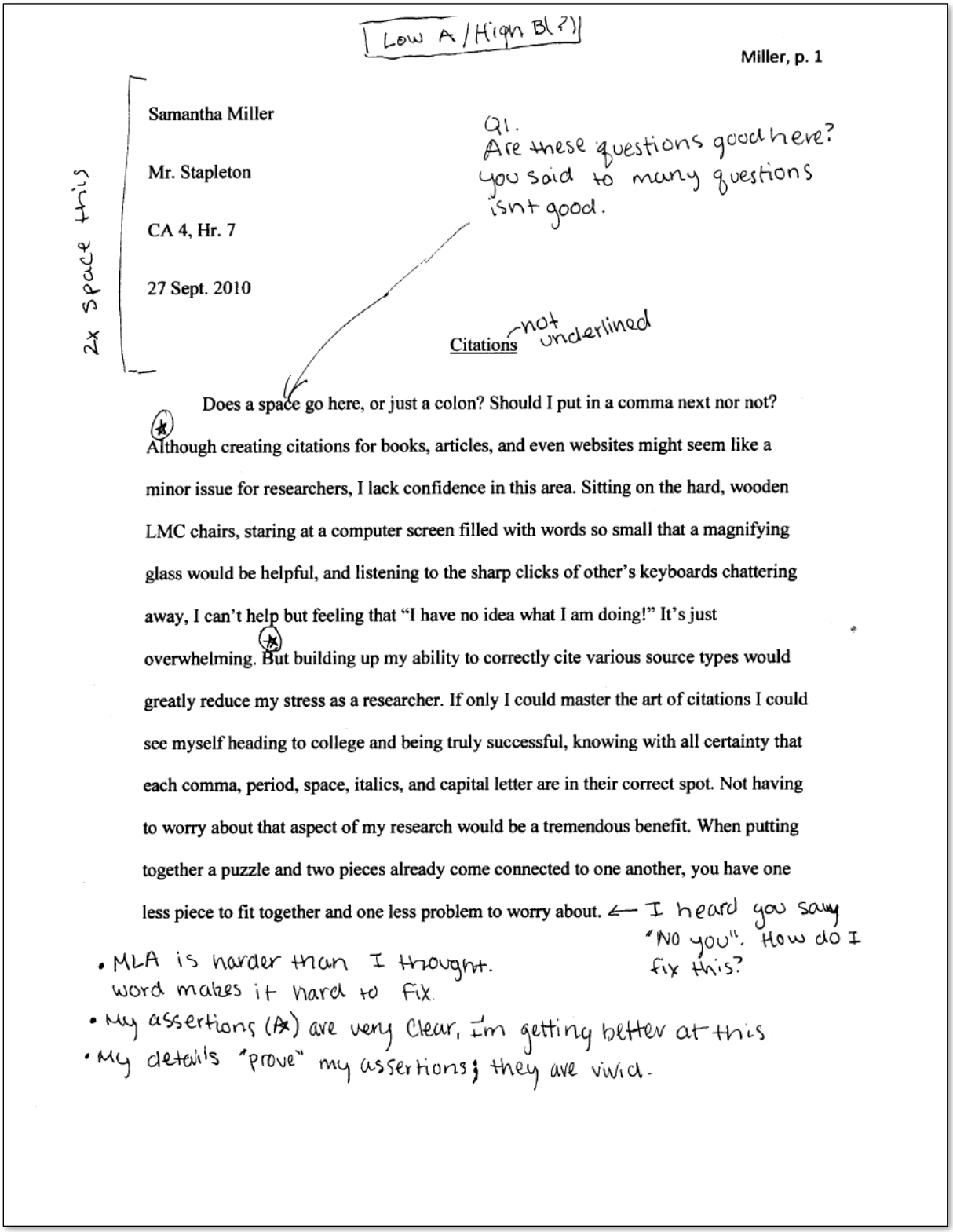
CA 4, Hr. 7

27 Sept. 2010

Citations

Does a space go here, or just a colon? Should I put in a comma next nor not? Although creating citations for books, articles, and even websites might seem like a minor issue for researchers, I lack confidence in this area. Sitting on the hard, wooden LMC chairs, staring at a computer screen filled with words so small that a magnifying glass would be helpful, and listening to the sharp clicks of other's keyboards chattering away, I can't help but feeling that "I have no idea what I am doing!" It's just overwhelming. But building up my ability to correctly cite various source types would greatly reduce my stress as a researcher. If only I could master the art of citations I could see myself heading to college and being truly successful, knowing with all certainty that each comma, period, space, italics, and capital letter are in their correct spot. Not having to worry about that aspect of my research would be a tremendous benefit. When putting together a puzzle and two pieces already come connected to one another, you have one less piece to fit together and one less problem to worry about.

Appendix F: Text *HSA*$_{annotated}$



Low A / High B(?)

Miller, p. 1

2x space this

Samantha Miller

Mr. Stapleton

CA 4, Hr. 7

27 Sept. 2010

Q1.
Are these questions good here? You said to many questions isn't good.

Citations — not underlined

Does a space go here, or just a colon? Should I put in a comma next nor not?

Although creating citations for books, articles, and even websites might seem like a minor issue for researchers, I lack confidence in this area. Sitting on the hard, wooden LMC chairs, staring at a computer screen filled with words so small that a magnifying glass would be helpful, and listening to the sharp clicks of other's keyboards chattering away, I can't help but feeling that "I have no idea what I am doing!" It's just overwhelming. But building up my ability to correctly cite various source types would greatly reduce my stress as a researcher. If only I could master the art of citations I could see myself heading to college and being truly successful, knowing with all certainty that each comma, period, space, italics, and capital letter are in their correct spot. Not having to worry about that aspect of my research would be a tremendous benefit. When putting together a puzzle and two pieces already come connected to one another, you have one less piece to fit together and one less problem to worry about. ← I heard you saw "No you". How do I fix this?

- MLA is harder than I thought. word makes it hard to fix.
- My assertions (A) are very clear, I'm getting better at this
- My details "prove" my assertions; they are vivid.

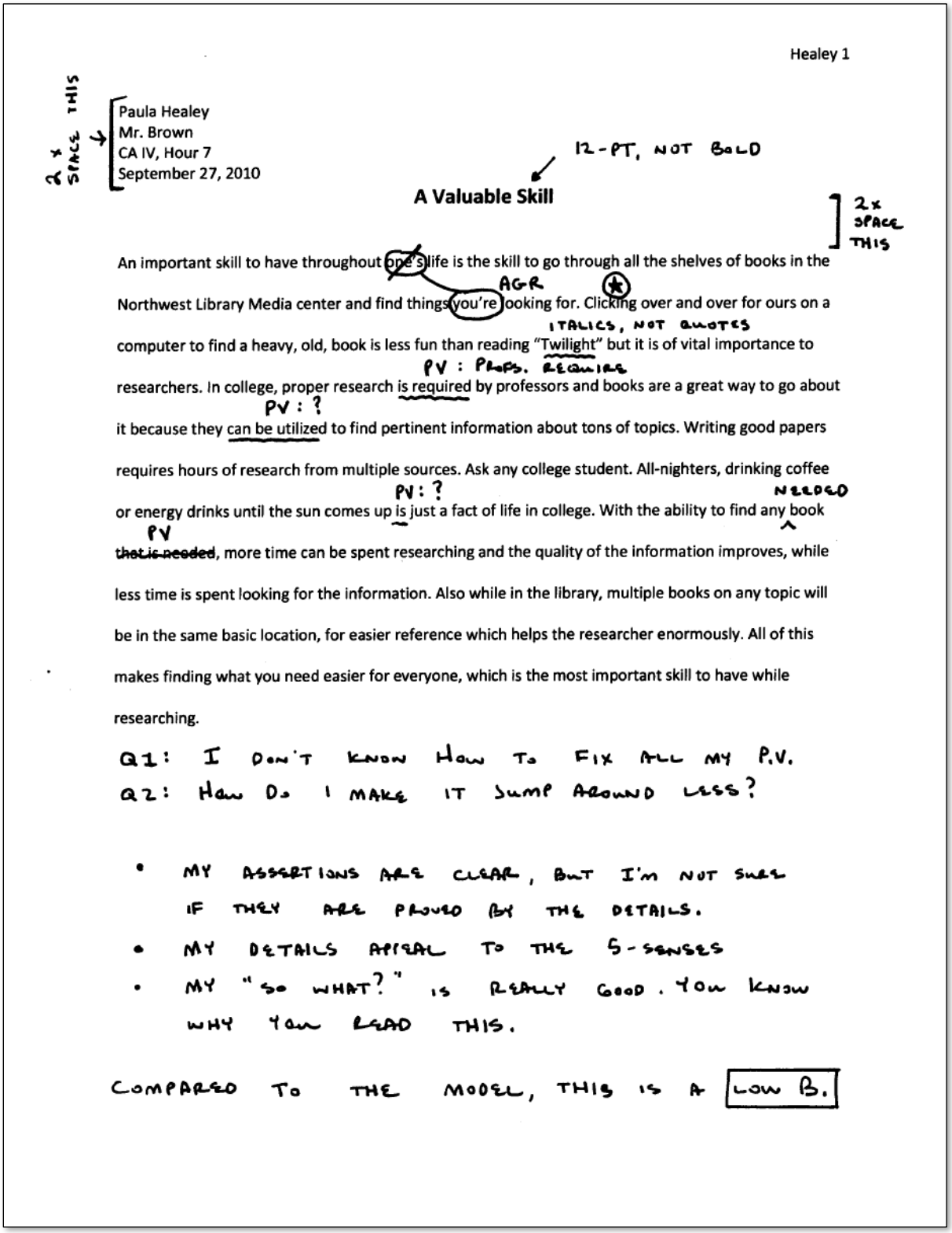
Appendix G: Text HSB_{clean}

Paula Healey
Mr. Brown
CA IV, Hour 7
September 27, 2010

A Valuable Skill

An important skill to have throughout one's life is the skill to go through all the shelves of books in the Northwest Library Media center and find things you're looking for. Clicking over and over for ours on a computer to find a heavy, old, book is less fun than reading "Twilight" but it is of vital importance to researchers. In college, proper research is required by professors and books are a great way to go about it because they can be utilized to find pertinent information about tons of topics. Writing good papers requires hours of research from multiple sources. Ask any college student. All-nighters, drinking coffee or energy drinks until the sun comes up is just a fact of life in college. With the ability to find any book that is needed, more time can be spent researching and the quality of the information improves, while less time is spent looking for the information. Also while in the library, multiple books on any topic will be in the same basic location, for easier reference which helps the researcher enormously. All of this makes finding what you need easier for everyone, which is the most important skill to have while researching.

Appendix H: Text *HSB*$_{annotated}$



Healey 1

2 x SPACE THIS

Paula Healey
Mr. Brown
CA IV, Hour 7
September 27, 2010

12-PT, NOT BOLD

A Valuable Skill

2x SPACE THIS

An important skill to have throughout ~~one's~~ life is the skill to go through all the shelves of books in the Northwest Library Media center and find things you're looking for. Clicking over and over for ours on a computer to find a heavy, old, book is less fun than reading "Twilight" but it is of vital importance to researchers. In college, proper research is required by professors and books are a great way to go about it because they can be utilized to find pertinent information about tons of topics. Writing good papers requires hours of research from multiple sources. Ask any college student. All-nighters, drinking coffee or energy drinks until the sun comes up is just a fact of life in college. With the ability to find any book ~~that is needed~~, more time can be spent researching and the quality of the information improves, while less time is spent looking for the information. Also while in the library, multiple books on any topic will be in the same basic location, for easier reference which helps the researcher enormously. All of this makes finding what you need easier for everyone, which is the most important skill to have while researching.

AGR

ITALICS, NOT QUOTES

PV : PROFS. REQUIRE

PV : ?

PV : ?

NEEDED

PV

Q1: I DON'T KNOW HOW TO FIX ALL MY P.V.
Q2: HOW DO I MAKE IT JUMP AROUND LESS?

- MY ASSERTIONS ARE CLEAR, BUT I'M NOT SURE IF THEY ARE PROVED BY THE DETAILS.
- MY DETAILS APPEAL TO THE 5-SENSES
- MY "SO WHAT?" IS REALLY GOOD. YOU KNOW WHY YOU READ THIS.

COMPARED TO THE MODEL, THIS IS A LOW B.

Appendix I: 8th-Grade Context, Prompt, and Performance Targets

8th-Grade Context and Prompt

It's mid September, and your eighth-graders are finally waking up from their summer doldrums. Rather than high-word count, low-frequency essays, they've been composing brief narratives on a bi-weekly basis. These shorter works allow for generous, high-value feedback in a timely manner, with only a few days between submission and return. The narratives also allow your students the opportunity to see, reflect upon, and revise (for additional credit) their work without feeling overwhelmed by mountains of corrections.

Recently, you have specifically been focusing their attention on the following writing-related skills:

- Narrative Writing: expressing a clear "purpose" through the use of an anecdote or brief story.
- Comma Usage: commas for introductory and interrupting elements, commas for incorporating dialogue.
- Paper Formatting:
    - Headings with the student's name, teacher's name, class, date.
    - Titles that are unique to the student's narrative, and reflecting its content.
    - Double-spaced text throughout.

For this essay—after reading a model story by Maya Angelou—you've asked your students to write a story about a specific lesson they've learned, using details and descriptions of the "teacher" and the setting of that lesson.

Performance Targets

Specifically, you are looking for the following traits in the paragraphs:

- Organization: A clear, interesting opening sentence that draws readers into the story without insulting them by a "this story is about" approach. Opening with a vivid quote or example is a good approach.
- Content, Organization: Clear progression through a "story" leading to (or informing) the purpose for which the story has been told—the meaningful lesson learned by the student.
- Content: Generous, vivid details to make the narrative's and purpose "spring to life."
- Content, Organization: A concluding sequence making the reader understand the story's importance.
- Sentence Fluency, Word Choice: Semi-formal language that is at once "personal" and authentic yet not overly casual for written, in-school communication. (*I*, *you*, and *contractions* are fine if effective).
- Conventions: Adherence to the conventions appropriate to personal/academic writing, allowing for logical/consistent departures where they enhance the work's effect.

Appendix J: 12th-Grade Context, Prompt, and Performance Targets

12th-Grade Context and Prompt

It's mid September, and your twelfth-graders are finally waking up from their summer doldrums. Rather than high-word count, low-frequency essays, they've been composing brief, one-paragraph pieces on a weekly basis. These shorter works allow for generous, high-value feedback in a timely manner, with only a few days between submission and return. The paragraphs also allow your students the opportunity to see, reflect upon, and revise (for additional credit) their work without feeling overwhelmed by mountains of corrections.

Meanwhile, your students have been pursuing three research-oriented goals in the library: (1) a review of the online catalog for finding books in the BVSD system; (2) an introduction to the databases for literary research; and (3) practice with MLA conventions for documenting print and electronic sources. They've done this work in preparation for their upcoming research papers in the second quarter.

Rather than giving quizzes or exams, you're assessing for understanding by way of performance tasks and reflective writings. For this reflective piece, you've asked students to explain in one paragraph either (a) what research skills they have learned best in the last five weeks of work or (b) what skills they still most need to learn before beginning the research paper in late October.

Performance Targets

Specifically, you are looking for the following traits in the paragraphs:

- Organization: A clear, interesting context-setting sentence alerting readers to the issue of this paragraph without insulting them by an obvious two- or three-prong statement.
- Organization, Content: One or two clear, organically coherent major assertions in the paragraph.
- Content: Generous, vivid details to make the paragraph's ideas "spring to life."
- Content: Commentary answering readers' questions about importance (i.e., "So what?" statement).
- Sentence Fluency, Word Choice: Semi-formal language that is at once "personal" and authentic yet not overly casual for academic/professional communication. (*I*, *you*, and *contractions* are fine if effective).
- Conventions: Adherence to MLA formatting standards; adherence to the conventions appropriate to personal/academic writing.

Appendix K: 8th-Grade Teacher Exemplar Text (1 of 2)

Bobby Blue Valley

Mr. Richardson

8th-Grade CA

15 September 2010

Lesson Learned

"Just be yourself. It's much better for you to tell us that you don't know than it is for you pretend you know something you don't." So said Professor Frey this past summer when I went in to his office to discuss my upcoming comprehensive exam. After we had talked about the eight-item list of possible questions—from which he would select three on exam day—I had asked him if he had any particular advice for preparing. After all, the questions were on one hand really simple but on the other hand impossible to ever answer fully—at least for a student like me.

All together, his suggestions were pretty solid all the way through. For example, he told me that the trick to not becoming overwhelmed was to have a specific audience in mind as I wrote each response. He even suggested what the audience should be. "You know," he said, "from what you've told me, it sounds like you're going to spend the rest of your career working with high school teachers rather than along an office row of statisticians. So it really isn't really important for you to explain the *Central Limit Theorem* or Cook and Campbell's threats to validity to someone like me. What you really need to do is be able to make these ideas spring to life and make sense for your real colleagues. Why don't you aim for that?"

And that's exactly what I did while reviewing for the exam, which itself was rather anticlimactic—taking place in a dusty, disused office, on a laptop computer that must have been at least ten years old. In fact (and how crazy is this?), every ten minutes of my four-hour, timed exam, the computer beeped and asked me if I wanted to install its new software "now" or be reminded "later." Somehow I managed to get through. And then something pretty magical

Appendix K: 8th-Grade Teacher Exemplar Text (2 of 2)

happened four weeks later when I faced Professor Frey for the oral defense of my written work. On that day, he asked maybe the scariest/most fun question I've ever met: "OK, Bobby, I'm really interested in only one question today: *What do you know well, and what do you merely know well enough in order to pass the exam*."

As he finished delivering his question, I was so happy that I had stayed just those few extra minutes during our summer meeting, because those minutes gave me perhaps my best answer of the day. "Professor Frey," I said, "Do you remember when you told me last month to just be myself? Well, it was what you said next that I'm going to use in answering this question. If I remember it right, you said, 'Keep in mind that we've all been through the process of comprehensive exams, too, and we all remember that "passing" didn't mean that our professors couldn't stump us. It was *years* after I finished my program before I felt like I had filled all my major gaps.' Well, on the heels of that, all I can say today is that I'm not sure I know any of it 'well.' What I have is a good tool box and some really good notes and reminders about how to use it. It will probably be years before I know any of it well. That's the honest truth."

His answer? *You passed*.

Appendix L: 12th-Grade Teacher Exemplar Text

Betty Blue Valley

Mrs. Bergstrom

CA IV, Hour 3

15 September 2010

Five-Sentence Summaries: Who Knew?

I remember from last year how hard it was to find twenty good sources to consider for my "betterment of the community" presentation. Actually, the problem wasn't so much *finding* the sources but remembering well enough what all the sources had to say so that I could decide which ones might best be synthesized into an organic ten-source packet. For example, until I had read several times George Count's statement in *Dare the School Build a New Social Order?* that "the school must shape attitudes, develop tastes, and even impose ideas,"[1] I didn't have the ability to see how it might fit together with Stephen J. Gould's warning that "that "[f]ew tragedies can be more extensive than the stunting of life, few injustices deeper than the denial of an opportunity to strive or even to hope, by a limit imposed from without, but falsely identified as lying within." If only I had known about five-sentence summaries, I could have more quickly put that packet together. Taking just three minutes at the end of reading each source to write down a quick overview of the source's argument and to add one or two of my favorite passages would have been enough in a first-pass reading to have shortened my selection process by half. How much easier is it, after all, to re-read and get a "global view" of twenty five-sentence summaries than twenty full articles? Thinking about having this skill in my back pocket, I'm almost *excited* about the research paper I'm going to have to write later in the spring.

---

[1] Teachers, please note that students were not *required* to use quotations in writing this paragraph. Further, because this was a fairly informal, reflective piece, there was no requirement for parenthetical documentation or a works cited list.

Appendix M: Control-Group Scoring and Feedback Instructions

Your Tasks as Evaluator

You must assign a grade. But in keeping with NCTE and National Writing Project wisdom, you are evaluating this essay in the hopes that students will use your comments to understand not only where the piece is now but also what may be done to improve its outcomes—including the grade in eSIS—should your students choose to do so.

IMPORTANT: Please give the paper the grade it actually deserves. Because you allow for rewrites, the grade doesn't have to reflect anything other than your honest standards.

- Give the paper **A PERCENTAGE GRADE** according to your usual interpretation of the following traditional scale: A (90-100%), B (80-89%), C (70-79%), D (60-69%), F (<60%).
- Make comments to **justify your grade**.
- Make comments to **help the student improve** on a subsequent draft of *this* piece.

Appendix N: Experimental-Group Scoring and Feedback Instructions

Your Tasks as Evaluator

You must assign a grade. But in keeping with NCTE and National Writing Project wisdom, you are evaluating this essay in the hopes that students will use your comments to understand not only where the piece is now but also what may be done to improve its outcomes—including the grade in eSIS—should your students choose to do so.

IMPORTANT: (1) Please give the paper the grade it actually deserves. Because you allow for rewrites, the grade doesn't have to reflect anything other than your honest standards. (2) *HOWEVER*, where the author of this paper has already made corrections to the essay or asked meaningful questions, feel free to credit the existing annotations to the student's grade (3) As you make comments, do not feel the need to repeat what the student has noted already.

- Give the paper **A PERCENTAGE GRADE** according to your usual interpretation of the following traditional scale: A (90-100%), B (80-89%), C (70-79%), D (60-69%), F (<60%).
- Make comments to **justify your grade**.
- Make comments to **help the student improve** on a subsequent draft of *this* piece.

Appendix O: Informed Consent (1 of 2)

INFORMED CONSENT STATEMENT

Improving Interrater Reliability and Quality of Feedback
in 8-12 English Language Arts Writing Assessments

INTRODUCTION
The School of Education at the University of Kansas supports the practice of protection for human subjects participating in research. The following information is provided for you to decide whether you wish to participate in the present study. You may refuse to sign this form and not participate in this study. You should be aware that even if you agree to participate, you are free to withdraw at any time. If you do withdraw from this study, it will not affect your relationship with this unit, the services it may provide to you, or the University of Kansas.

PURPOSE OF THE STUDY
This study seeks to gather the following types of information from experienced English Language Arts teachers regarding their current practices in evaluating written work and/or how these practices might change under subtly different conditions for instruction and assessment.

(a) The grades teachers attach to a pair of sample essays simulating work done by 8th- or 12th-grade students.
(b) The quantity and characteristics of feedback given to these essays under a realistic pressure of time.

PROCEDURES
For a fifteen-minute time period, you will be asked to provide authentic responses to two sample essays. The first part of your response task will ask you to assign grades according to a modified version of the 6-trait rubric, which breaks essays up analytically according to the following characteristics: *content and ideas*, *organization*, *sentence fluency*, *word choice*, *conventions*, and *voice*. The second part of your response task will ask that you provide written feedback you would consider useful in helping the writer understand the essay grade and/or improve the essay's outcomes in a subsequent draft. Following the fifteen-minute time period, we will use the data you have provided as a springboard into a 45-minute conversation about current assessment practices in our discipline. The data will be kept in a locked cabinet during the course of my study; afterwards they will be removed to a secure, off-site location.

Unless you specifically choose otherwise, your work will be received and treated as anonymous data. It will only be seen by the principal instructor (Charles Golden) and an non-administrator assistant trained to classify feedback responses according various traits to be discussed during the professional development session. Following the completion of the research project, the data will be stored in a secured, off-site location should this study lead to further related research.

RISKS
There are no foreseeable physical or professional risks in completing the research task. None of your responses will be reported to district personnel. Any data presented by way of conference sessions, dissertation materials, or published documents will not capable of being traced to any particular individual within the district.

BENEFITS
My professional/academic hope in completing this research is to engage BVSD communication arts teachers in an ongoing dialogue towards better writing outcomes for all our students. At minimum, this project will help focus upon one possible way forward toward such outcomes.

PAYMENT TO PARTICIPANTS
You will not be compensated in any way for participation in this study.

Approved by the Human Subjects Committee University of Kansas, Lawrence Campus (HSCL). Approval expires one year from 10/9/2010. HSCL #18975

Appendix O: Informed Consent (2 of 2)

PARTICIPANT CONFIDENTIALITY

Unless you choose otherwise, your name will not be associated with the information collected about you or with the research findings of this study. Moreover, I will share no information about your individually identifiable responses unless you have approved a for-publication or for-presentation draft of my use in advance.

There is no expiration date to the information used in this study. By signing this form you give permission for the use and disclosure of your information for purposes of this study at any time in the future.

REFUSAL TO SIGN CONSENT AND AUTHORIZATION

You are not required to sign this Consent and Authorization form and you may refuse to do so without affecting your right to any services you are receiving or may receive from the University of Kansas or to participate in any programs or events of the University of Kansas. However, if you refuse to sign, you cannot participate in this study.

CANCELLING THIS CONSENT AND AUTHORIZATION

You may withdraw your consent to participate in this study at any time. You also have the right to cancel your permission to use and disclose information collected about you, in writing, at any time, by sending your written request to: Charles H. Golden, Blue Valley Northwest High School, 13260 Switzer, Overland Park, KS 66213. If you cancel permission to use your information, the researchers will stop processing the data you have provided. However, the research team may use and disclose information that was gathered before they received your cancellation, as described above.

QUESTIONS ABOUT PARTICIPATION

Questions about procedures should be directed to the researcher(s) listed at the end of this consent form.

PARTICIPANT CERTIFICATION:

I have read this Consent and Authorization form. I have had the opportunity to ask, and I have received answers to, any questions I had regarding the study. I understand that if I have any additional questions about my rights as a research participant, I may call (785) 864-7429 or (785) 864-7385, write the Human Subjects Committee Lawrence Campus (HSCL), University of Kansas, 2385 Irving Hill Road, Lawrence, Kansas 66045-7568, or email mdenning@ku.edu.

I agree to take part in this study as a research participant. By my signature I affirm that I am at least 18 years old and that I have received a copy of this Consent and Authorization form.

_____ _____

Type/Print Participant's Name Date

_____

Participant's Signature

Researcher Contact Information

Charles H. Golden
Principal Investigator
Blue Valley NW HS
13260 Switzer
Overland Park, KS 66213
(913) 239-3400

Heidi Hallman, Assistant Professor
Curriculum & Instruction
University of Kansas
Joseph R. Pearson Hall, 338
1122 W. Campus Rd.
Lawrence, KS 66045
(785) 864-9670