Journal of Business and Technical Communication
Volume 23 Number 1
January 2009 83-107

10.1177/1050651908324383
http://jbtc.sagepub.com
hosted at
http://online.sagepub.com

Methods and Results of an Accreditation-Driven Writing Assessment in a Business College

Scott Warnock
Drexel University

This article describes a pilot effort for an accreditation-driven writing assessment in a business college, detailing the pilot's logistics and methods. Supported by rubric software and a philosophy of "real readers, real documents," the assessment was piloted in summer 2006 with five evaluators who were English instructors and four who worked or taught in business environments. The nine evaluators were each given 10 reports that were drawn from a sample of 50 reports completed in a writing-intensive course. They created 88 individual assessments using a 10-category rubric. While the overarching purpose of the pilot was to determine the effectiveness of the methods used, the results may also be of interest to those involved with the assessment of writing.

Keywords: *accreditation; business writing; rubrics; writing assessment*

Accreditation, as Yancey and Huot (1999) pointed out, can provide the needed "exigence" for assessment (p. 12). Therefore, in response to the accreditation requirements of the Association to Advance Collegiate Schools of Business (AACSB), faculty in Drexel University's LeBow College of Business identified four learning goals to assess: problem solving, knowledge of economics, career learning, and writing skills. For each of these, faculty needed to devise methods for the ongoing assessment of student performance as well as assessments of curricular and instructional interventions aimed at improving performance on a continual basis.

LeBow faculty included writing because they recognized, as LeBow associate dean F. Linnehan (personal communication, February 27, 2006) commented, "the critical importance of written communication in business today" as communicated to them by employers. But LeBow needed a

Author's Note: I would like to thank Frank Linnehan and Peter Heisen for their help in gathering data and writing this article.

method of measuring its students' writing competency and performance. Large-scale assessment in education has proven to be notoriously difficult, particularly in subjective areas such as writing. Although large educational studies have been undertaken (e.g., Applebee, Langer, Nystrand, & Gamoran, 2003; Haswell, 2000; Langer, 2001; Sommers, 2002; Xue & Meisels, 2004), including in business and technical writing settings (Zhao & Alexander, 2004), correlating specific shifts in pedagogical strategies and techniques with changes in student performance has been challenging. For instance, Haswell (2000) reported that in 1963, Kitzhaber started the project of documenting how college students change their writing performance, but "since then, it is only fair to say the project has not been much advanced by researchers" (p. 307). And large overviews of student writing have been difficult to undertake because of the complexity and many constraints inherent in such a project.

Faced with the challenge of assessing writing, LeBow initiated a collaboration with the Department of English and Philosophy at Drexel. Several LeBow faculty members met with members of the English department in spring 2005 to describe the accreditation process and how writing would be a component of it. Working together, these two groups recognized that an accreditation-driven assessment could do more than simply satisfy AACSB: An extensive writing assessment could be the impetus for launching a novel process—one that would ideally be reproducible—to enhance the writing instruction at LeBow. Conducted properly, the accreditation-driven assessment of LeBow undergraduates' writing could be "an opportunity to learn something worth learning" (Yancey & Huot, 1999, p. 8). As Williamson (1999) suggested, "Even a mandated external evaluation can be co-opted by a department that is willing to accept the invitation to examine itself with an eye to improving its understanding of its program" (p. 254). Of course, others have recognized the problems I mention here. Huot (1999) cited Berlak in cautioning that "historically, educational assessment has demonstrated the potential to produce results that satisfy outside pressures for accountability without producing any information of value to the programs themselves" (p. 69). So we had to be sure that, in embracing accreditation, we created an assessment with both local and global outcomes.

This article has two purposes. The first and primary purpose is to describe the methods used to pilot this assessment in summer 2006 and their theoretical underpinnings. Although pushed by the AACSB, LeBow faculty realized that the assessment could (a) demonstrate the strengths and weaknesses of student writing and (b) establish grounds for changes in the curriculum to address these findings. The objective of the pilot was to determine the effectiveness of this

assessment and to troubleshoot problems, loosely in the spirit of a phase I clinical trial. But after we evaluated the statistical results from the pilot, we discovered that despite the small sample size, several aspects of the results could be of interest to those involved with writing assessment—including the way that the results show how different groups of evaluators respond to the same documents. So, the second purpose of this article is to provide those results and an accompanying commentary, which includes how the data might help frame future iterations of this assessment.

Methods for the Pilot

One of the first tasks that the LeBow collaboration with the Department of English and Philosophy accomplished was creating a committee to guide the assessment process. The LeBow Writing Committee consists of faculty and staff from the School of Business and the Department of English and Philosophy as well as LeBow alumni. I am the head of that committee and the coordinator of the summer pilot. One of the committee's first tasks was to set up the summer assessment described here. Because the ultimate goal of the assessment is to look broadly at the writing that takes place in LeBow in order to "address the total nature of the program rather than its individual components" (Yancey & Huot, 1999, p. 8), we decided to look at students' writing from an actual LeBow classroom assignment rather than administer a timed examination or something of that nature. As Huot (1999) articulated, "It is more sensible to use writing being written by students in that program, rather than using writing samples produced specifically for evaluation" (p. 72). For this pilot, we randomly selected 50 samples from a large sample of student papers written for 3rd-year courses in organizational behavior, a writing-intensive course taken by all LeBow students. We removed all student identifiers and coded the samples.

The assignment, a three- to eight-page report, had two parts: In the first part, the students were asked to choose 10 concepts from their organizational behavior text

> and provide examples that clearly illustrate each one. The concepts you select must be chosen from the key words that are highlighted throughout the chapters of the text. Your examples can be real or made up; they may be based on work or non-work experiences, but cannot be from the text or from our class discussions.

They were to devote about two pages to this part of the assignment. In the second part, the manager interview, students were to provide the results of

an interview they conducted with a manager of a business. The assignment focused on applying principles of organizational behavior to the practices of working professionals; students were given a guideline of questions to pose to their interviewees, questions intended to help students contemplate how the principles of the course applied to these managers' everyday decisions. They also needed to prepare questions of their own. The assignment instructions told students that they would be graded on the quality of their questions, their description of the interviewee's answers, their analysis of the interview, their ability to connect the interview to conceptual material from class and the text, and their writing (based on a rubric that was similar to the one used for our assessment). Because the goal of the pilot was to test the assessment process, this single group of reports was satisfactory. Future iterations of the assessment will involve writing from a broader selection of writing-intensive courses.

The committee recruited 10 evaluators: 5 were English faculty and 5 were businesspeople (3 Drexel graduates with positions higher than entry level and 2 recent MBA graduates). One of the business evaluators dropped out of the assessment, leaving 9 evaluators. We assigned each evaluator an identification code and asked each to assess 10 reports. For the pilot, we used a straightforward method to assign the reports to the evaluators: They each shared five reports with 2 other evaluators (e.g., Evaluator 2 shared reports 1 through 5 with Evaluator 1 and reports 6 through 10 with Evaluator 3). So each report was assessed by 2 evaluators. We will use a more elaborate randomization scheme in full-fledged assessments in upcoming summers, but for the purposes of the pilot, for which uncovering flaws in the process was paramount, this method adequately provided a measure of variety (as opposed, e.g., to simply pairing evaluators and having them read the same 10 reports). To meet accreditation requirements, LeBow established a budget for the assessment that included a $100 stipend for each evaluator.

The committee then developed a set of criteria that evaluators could use to assess the reports. I created a rubric with those criteria and distributed it for feedback to LeBow faculty, who asked questions and provided useful comments about the rubric, in some cases, based on their use of it in their courses. In creating this rubric, I considered that "individual guidelines need to be articulated for specific disciplines. It should be apparent that only those who teach, write, and work in specific areas are really competent to set guidelines and evaluate student writing" (Huot, 1999, p. 72). Faculty commented by contacting members of the committee or during a meeting to discuss the assessment initiative.

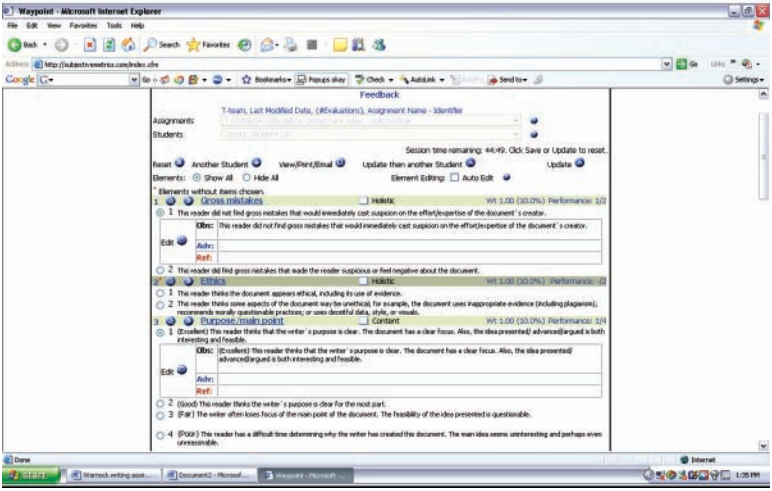
Figure 1
The Ten Rubric Elements

Elements and Scoring Criteria

Element Name	*Description of Purpose*	*Reviewer Choices and Scoring for Assessment*
Gross mistakes	Did the reader find gross mistakes/errors in the document?	No = 1.0 Yes = 0.0
Ethics	Did the reader think that some aspect of the document was unethical?	
Purpose/main point	How clear is the purpose of the document?	Excellent = 1.0 Good = 0.67 Fair = 0.33 Poor = 0.0
Audience	Is the audience of the document clear?	
Organization	Is the document organized clearly?	
Evidence	Does the document use evidence effectively?	
Sentence style: Flow of writing	Is the writing in the document clear? Does the writing flow well?	
Correctness: Grammar and writing mechanics	Is the document grammatically and mechanically correct?	
Document design/appearance	Is the document designed effectively?	
Visuals if applicable (tables, charts, pictures, etc.)	Does the document utilize visual elements effectively?	

After receiving this feedback, I revised the rubric and loaded it into a Web-based assessment tool, Waypoint (a product of Subjective Metrics, a company for which I am a founder and minority shareholder). In brief, Waypoint provided a means for the evaluators to assess the reports. Figure 1 shows the 10 rubric categories, or elements, and Figure 2 shows a screen capture of how a portion of the rubric appeared in Waypoint (to read the complete rubric, see the appendix). The evaluators used the rubric to assess each of their 10 reports. As Figure 1 shows, two of the rubric's elements, gross mistakes and ethics, are binary. Our rationale for including the two

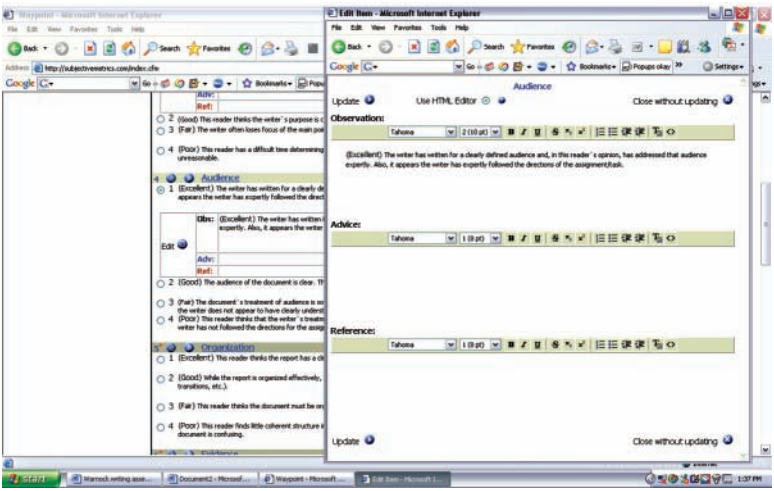
Figure 2
Screen Capture of the Rubric in Waypoint



binary elements was to determine a kind of knee-jerk reaction that a reader might have to writing that evidences some type of ethical issue, such as plagiarism, or overt mistakes, such as the misspelling of a client's name; for these two categories, then, the evaluator's assessment choice is either *yes* or *no.* The evaluators used a Likert-type 4-point scale, ranging from *excellent* to *poor*, to score the other eight elements. These four-item assessment choices reflect a way of thinking through a scale for different standard writing criteria. As Figure 2 shows, accompanying each assessment choice is language designed to help guide the evaluator. When the evaluators made a choice in Waypoint, they could open a pop-up window that allowed them to write additional comments if desired; Figure 3 shows the pop-up window for an evaluator who chose *excellent* for the audience element.

Although Waypoint gave the evaluators a unified, Web-based way to respond to the reports, it also served another purpose for those of us conducting the evaluation. As Figure 1 shows, each assessment choice correlates to a numerical value. Waypoint records all evaluator comments as well as the numerical value of the assessment choices for each element. It calculates an overall report score automatically based on these numerical

Figure 3
Pop-Up Window For the *Excellent* Choice in the Audience Element



values, using a straightforward algorithm. (Similarly, for an instructor using Waypoint, the rubric could generate an overall grade based on the average report score.) Thus, Waypoint provided a simple method of assembling our assessment data; all the data that I describe here were generated in this way.

In early summer 2006, we conducted a 90-minute training session for the evaluators. The focus of this session was on training the evaluators to use Waypoint, and we provided them with their Waypoint login information. We recorded the session and made it available online via a Web link. Several evaluators attended the session; others viewed the video online. After training, the evaluators received their 10 reports via e-mail. They had 1 month to assess their 10 reports (using Waypoint).

Theoretical Underpinnings of These Methods

The evaluators read the documents and assessed them without using holistic scoring strategies. During the training, we spent most of the time making sure the evaluators understood the software. We also discussed our writing rubric so that the evaluators would become comfortable with it. But we

stressed the importance of their own responses to the writing, explaining that we wanted their authentic responses as "real" readers of these real documents.

Although we did not go into great detail during the training, the concept behind our approach is that what stifles many writing assessment efforts, logistically speaking, is an illusionary game of validity in which evaluators struggle to establish idealized versions of the writing traits that meet specified writing criteria. In a book review, Whithaus (2005a) commented that concepts of validity and reliability emerged

> in the field of psychological testing and measurement and then continued to be applied to large-scale writing assessment even as test developers and composition researchers noted "that reliability and validity are troublesome criteria, inadequate and too limited for the distinctive task of evaluating writing." (p. 215)

We created an assessment structure in which the subjectivity of readers' natural responses to writing is an asset to the assessment rather than a problem that we would need to work around through norming or holistic means. We embraced the idea that, as C. Whithaus (personal communication, November 15, 2007) put it, our readers are "encouraged to hold onto their 'situatedness' rather than to hide that" behind conventional norms. The reception of any writing, of course, depends on variables that are difficult, if not impossible, to quantify. We based our approach, however, on the idea that a large enough group of real readers can assess writing in a way that makes a useful, representative statement about the way writing will be received by audience members for whom it may pertain.

An obstacle in writing assessment has been in creating reasonable structures that allow large groups of assessors access to large numbers of documents. The methods of our pilot assessment along with their theoretical underpinnings may provide a reproducible model that allows for such, thus expanding the opportunities for schools, departments, and programs to perform meaningful writing assessments. This approach may help writing researchers address some difficulties brought about by the inherently subjective practice of assessing writing. Although eschewing norming strategies may certainly cause problems in an assessment with a small number of evaluators, with larger numbers, individual subjectivity is elided. Consider, as an analogy, the use of a postoperative pain scale in orthopedics. Of course, pain is subjective, so if Patient A says that implant X hurts an 8 on a scale of 1 to 10 and Patient B says that implant X hurts a 4, we know little. But when hundreds, maybe thousands, of patients make assessments about implant X on that same scale, statistical power flattens individual differences, allowing researchers to make a statement about the pain factor of

implant X. In like manner, if we can create an assessment structure that can accommodate large numbers of evaluators and documents but is simple enough to be reproduced, then we will have a ready tool for large-scale writing assessments, not just a specialist's method of evaluating writing. By using large numbers of assessors and reports, we can gain a full picture of writing and yet still recognize the inherently, unavoidably subjective views of any writing evaluator.

This approach also gave us an opportunity: At the same time that we were assessing student writing, we were also gaining insight into potential differences in the way writing is viewed by English instructors versus how it is viewed by businesspeople. We believed the results would reveal interesting differences in audience and, perhaps, suggest how those differences might affect writing in business schools. For instance, if students receive one type of assessment in courses such as 1st-year writing and business writing, courses often designed and taught by members of the English department, and another from writing-intensive courses in their major, would students be confused by those differences? A further, more ambitious study of these audience differences might reveal how our expectations and instruction of writing are influenced by disciplinary backgrounds. This article does not address these issues, but as this assessment continues, perhaps answers will emerge.

Results of the Evaluators

In the pilot, the goal was that the 10 evaluators would each assess 10 reports, for a total of 100 assessments. But, as I mentioned, 1 business evaluator withdrew from the pilot. The evaluators reported few problems operating Waypoint, but 2 evaluators had trouble opening the Word file of one of the reports. So 7 of the evaluators assessed all 10 of their reports, and 2 evaluators assessed 9 reports, for a total of 88 assessments.

Of the 50 reports, 17 were evaluated by both an English instructor and a businessperson, 16 reports were evaluated by two businesspeople, and 17 were evaluated by two English instructors. Again, we designed the distribution method to uncover flaws in the assessment process, but after compiling the data, we noticed some interesting differences between the business and the English evaluators. In the future, we will more systematically distribute the documents to further investigate these differences.

In what follows, *overall score* refers to the score calculated by Waypoint for an assessment (almost like the grade an instructor might assign a given report based on the evaluation choices), which is based on the average of

Figure 4
Overall Scores for Reports of All Evaluators (N = 88)

the element scores. *Element score* refers to the score of a particular element. Values for both overall scores and element scores are reported on a scale from 0.0 to 1.0. Again, most of the elements had 4-point scales, and for those elements, the scoring was 1.0 for *excellent*, 0.67 for *good*, 0.33 for *fair*, and 0.0 for *poor* (see Figure 1). On the 2-point scales for ethics and gross mistakes, reports received either a score of 0.0 (yes) or 1.0 (no): The report either was or was not unethical, and it either had or did not have gross mistakes. Note that the assignment used for the pilot did not invite the use of visuals, so the visuals element was infrequently scored, and no results from this element are included in the data reported here.

Overall Scores

The graphs in Figures 4, 5, and 6 demonstrate the overall scores for the reports of all the evaluators, the overall scores of the business evaluators, and the overall scores of the English evaluators. The evaluators gave a perfect score of 1.0 nine times. Thirty-four of the reports (39%) fell between 0.7 and 0.8, and 59 reports (67%) received an overall score of 0.7 or greater (i.e., a score of *good*; see Figure 4). The mean overall score of business evaluators was 0.62 (see Figure 5) whereas the mean overall score for the English evaluators was 0.77 (see Figure 6). This difference is statistically significant (t test, $p < .001$).

Figure 7 shows the interesting difference in many of the overall scores of the 17 reports (as identified by the student identification code for the evaluation)

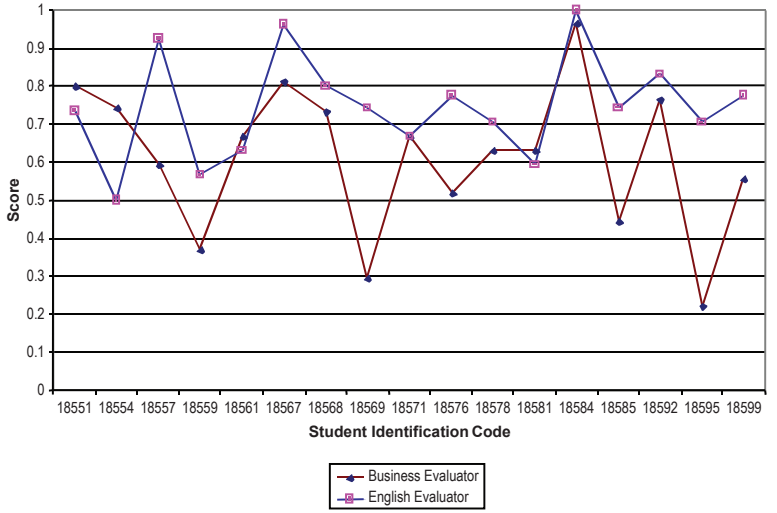
Figure 5
Overall Scores for Reports of Business Evaluators (n = 38)

Figure 6
Overall Scores for Reports of English Evaluators (n = 50)

that were assessed by both a business evaluator and an English evaluator. For these 17 pairs of assessments, the overall score of the business evaluator was higher than that of the English evaluator in only four instances.

The average overall score for each evaluator is shown in Figure 8. Of the nine evaluators (identified only by their identification codes), the four who gave the highest scores were English instructors, and the three who gave the

Figure 7
Overall Scores on Reports Evaluated by Both a Business Evaluator and an English Evaluator ($n = 17$)

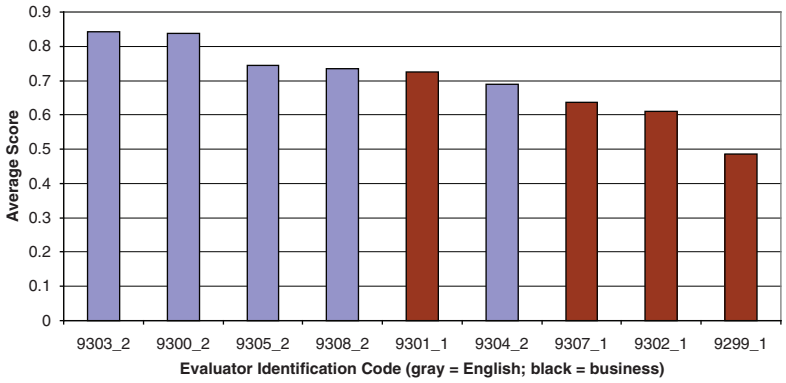


lowest scores were businesspeople. In fact, except for a transposition at the midpoint between two evaluators who were close in their scoring (i.e., the difference in their average overall scores was statistically insignificant), all the English evaluators scored the reports on average higher than did all of the business evaluators.

Element Scores

Figure 9 shows the average element scores for the evaluators overall and for each of the two groups. Overall, the evaluators' highest average element scores were in the binary categories of ethics (0.98) and gross mistakes (0.89) because most of the evaluators judged the reports as being ethically sound and free of gross errors. Of the overall average scores for the 4-point elements, purpose or main point (0.72) and document design (0.72) were highest, followed by organization (0.67), audience (0.65), evidence (0.58), and sentence style (0.57) and correctness (0.57). Although I acknowledge

Figure 8
Average Overall Score for Each Evaluator (n = 9)

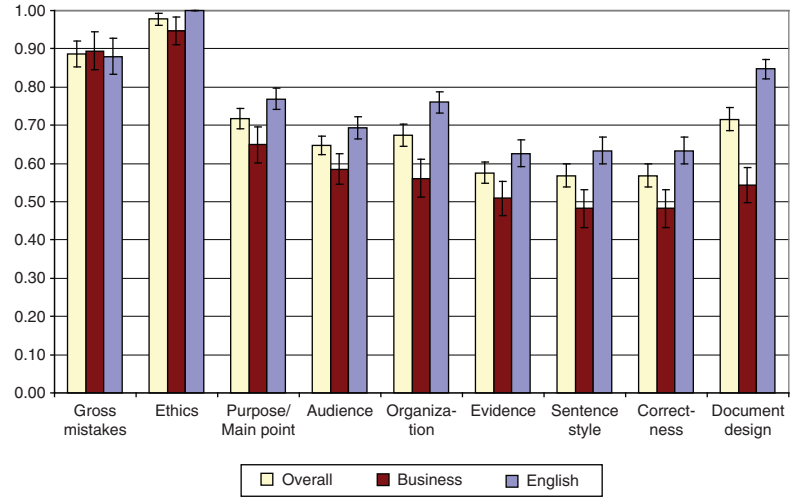


that the numerical scale we established is arbitrary (e.g., in correlating an assessment of *good* with the number 0.67), the overall scores for four of these elements were below 0.67, meaning the evaluators in general viewed them as less than good. In comparing the average element scores between the two groups of evaluators, for every element except gross mistakes (one of the binary elements), the business evaluators scored the reports lower than the English evaluators did. More important, for all of the 4-point elements, the English evaluators ranked the reports significantly higher (i.e., the error bars do not overlap) than the business evaluators did.

Evaluators' Edited Commentary

Although the evaluators used the rubric as a guide, they were able to make additional comments by clicking the Edit button (see Figure 3). Such comments become part of the assessment data and can be retrieved by those administering the assessment. Out of more than 800 opportunities for commentary, the evaluators added comments on only 31 individual elements; most evaluators allowed the standard language of the rubric to capture their assessment of the reports. Only in two cases did evaluators—both businesspeople—identify reports as having ethical lapses (i.e., they gave those reports a score of 0 for that binary element), and in each of those cases, the

Figure 9
Average Element Scores



evaluator wrote a clarifying comment. In one case, the evaluator thought that the person whom the student interviewed for the report was related to the student. In the other case, the evaluator wrote the following:

> The content of this paper was unethical, but not because of the student's opinion. The interviewee had some unethical practices that were shared in the paper. I thought the student [should have] recognized this and commented more on the unethical nature of the interviewee's opinions and views.

The distinction between these comments seems important in investigating how these evaluators viewed the concept of ethics. In the first comment, the evaluator noted that the student who wrote the report did something unethical. In the second comment, the evaluator noted that the student who wrote the report did nothing ethically wrong but failed to recognize or emphasize the unethical behavior of the person whom the student interviewed for the report. Several evaluators from both evaluator groups who identified gross mistakes also chose to provide comments, such as these:

- The document is incomplete. No interview attached.
- This student needs major editing practice. It is hard to evaluate the subject matter of the content when there are so many obvious errors.
- I think this student has the wrong definition for Type A personality.
- The name of the document indicated some kind of gross mistake [which the evaluator did not explain further].

Having the ability to make such comments in a writing assessment allows evaluators, when necessary, to go beyond the checklist of a rubric scale. In contrast to a straight holistic scale, this method in which evaluators can provide easily stored and retrieved comments fits into Whithaus's (2005b) idea about how an assessment might work:

> If individuals are encouraged to record their responses, and these responses are then associated with the compositions in a database that is available to the instructor and evaluators, it is possible to build a situated evaluation of a student's composition. (p. 88)

Of course, in providing these comments, the evaluators may reveal something about their own writing skills, which could influence the way their evaluations are viewed. For example, an evaluator's comments may exhibit grammar or mechanical errors. If evaluators, then, reveal through their comments that they lack expertise in terms of grammar and mechanics, how, if at all, might that influence the way we value their assessments?

Discussion

Based on the success of our pilot assessment, we are prepared to move forward with larger (more than 300 documents) assessments at LeBow. By following these practices and concepts from the pilot assessment, we may be able to circumvent some of the problems with large writing assessments.

Avoid norming or holistic scoring practices. The value of norming is questionable, yet time, energy, and money are often spent on efforts to normalize readers by employing holistic scoring strategies that seek a "true score" for writing (Elbow, 1996, p. 85). Perhaps norming practices are useful for placement, but for other writing assessments, we agree with those who find norming to be an entrée to various problems because it sets up a fictional, idealized context in which to assess writing. In his critique of these practices, Huot (2002) argued that "conventional writing assessment's emphasis on uniformity and test-type conditions are a product of a testing

theory that assumes that individual matters of context and rhetoric are factors to be overcome" (p. 85). Earlier, Huot (1996) also discussed revising concepts of validity and testing theory to "strengthen . . . our ability to devise new and more appropriate measures for the evaluation of literate activity" (p. 114). And Whithaus (2005a) acknowledged that

> the usefulness of inter-rater reliability is on the way out in theory, and may even be changing in practice; validity is changing into a concept that not only can but must acknowledge the situation in which a given composition or series of compositions are produced. (p. 217)

Too often, it seems, writing-assessment research is handcuffed by exaggerated needs to normalize or synchronize the views of assessors; we believed that we could achieve meaningful results without ignoring the effects of context and by respecting the natural subjectivity of the task.

Overcome the obstacle of audience. We based our assessment on the concept that evaluators who are reasonably qualified readers and writers of English can, when guided by a rubric, make legitimate subjective decisions about a given piece of writing. After all, that is the way writing is typically judged—even in school. Further, those decisions can be used in the aggregate to draw conclusions about a sample of writing. Assessment is often stymied when those administering the assessment overthink the writing constraints. Huot (2002) suggested that

> an appropriate way to harness this tension is to base assessment practices within specific contexts, so that raters are forced to make practical, pedagogical, programmatic, and interpretive judgments without having to define writing quality or other abstract values which end up tapping influences beyond the raters or test administrators' control. (p. 102)

The spirit of Huot's ideas drives our concept of assessment based on the authentic responses of real readers in real contexts. Because the ultimate assessment of writing comes from audiences that by their very nature are varied, our design allows us—and, ideally, others—to view the different ways in which different audiences view writing and to incorporate those varied perceptions into the final results of our assessment.

Generate large numbers. According to the literature on writing assessment, large-scale assessments of writing are difficult to perform. They are sometimes developed in contrived situations—timed examinations, reviews of grades—that are rightly considered problematic by those interested in

assessment. Again, we can compare this work with the use of the pain scale in orthopedics, in which the statistical power generated by large numbers reduces individual subjectivity. This comparison is imperfect but useful because pain scales have been used to assess how large numbers of people make subjective decisions (hopefully the comparison between patients' rating of pain and evaluators' rating of student writing breaks down from there). A pain scale is not the only method of judging the value of a prosthetic, and we are not suggesting that our method is the sole way to assess writing at LeBow or elsewhere. What our method does, though, is provide a potentially reproducible method for evaluating large numbers of student writing samples. Generating large numbers of assessments elides outliers such as the curmudgeon, the grammar guru, or even the apathetic reader. The goal, then, is to create a structure that is flexible enough to generate large numbers of assessment records that statistically flatten out anomalies.

Although our approach to assessment might seem laissez-faire, its structure is hardly freewheeling. In the absence of norming, it produces many evaluations and thus a large body of data, a goal that composition research does not always seek. McLeod (2003) commented that because many compositionists emerge from the humanities tradition, they "tend to find the social-science world view an alien one—we like and do words, not numbers." Many even have "data discomfort." But, she added, "I have been distressed by the tendency in our profession to dismiss empirical research as based on positivist assumptions and therefore at best not contributing anything worthwhile, at worst something evil to be avoided." She suggested that "if we really celebrate diversity, we should extend that to our research paradigms" (pp. 152-153). As Taylor (2003) put it, "It may be more the case that the health of today's academic disciplines actually *requires* methodological diversity and interdisciplinarity rather than rigidity and insularity—much like a wide gene pool promotes immunity" (p. 145). Although we attempted to take a novel approach both in terms of method and philosophy, our overarching purpose was still to create a reproducible method of writing assessment.

For our local uses, the pilot was successful for several reasons. First, evaluators were able to use the assessment rubric with minimal training. The procedures and work flow were efficient and could easily be scaled up to accommodate a larger assessment. In addition, although we did not plan to evaluate the data, we obtained statistically significant results in comparing the two groups of evaluators—and from a relatively small sample. Going forward, we will conduct an assessment with 30 evaluators and 300 reports.

Those data—600 or more assessment records—will represent a large baseline sample drawn from several courses across the curriculum. We could most likely use these results to draw conclusions about the writing of LeBow students; for instance, instructors may be loath to change teaching patterns based on perceptions, but if our data show that students need to improve specific areas, such as use of evidence, then we could focus our faculty development efforts on those specific areas.

Again, our primary objectives for the ongoing assessment will be to evaluate the writing of LeBow students and to use the results to develop interventions aimed at continuous improvement. The average overall score for all the reports was 0.7, and the reports received an overall score of 0.7 or greater in 59 of the 88 assessments. Thus, the evaluators assessed the reports as being better than good (with *good* being a score of 0.67) in 67% of the assessment records.

At the element level, evaluators considered reports to be not ethically sound in only two cases and most to be free of gross mistakes; again, both of these elements were on a 2-point (yes or no) scale whereas the other elements were on a 4-point scale. Evaluators also indicated that the purpose or main point and document design or appearance of the reports were generally better than good. The average report scores, however, were below good for organization, audience, evidence, sentence style, and correctness, with the last three scoring below 0.60. If these scores were loosely correlated with a grade, students would receive about a C+ in these categories. And if our full-scale evaluation mirrors these results, then LeBow might develop ways to teach these three areas more rigorously across its curriculum. The data could help determine which areas need more attention.

Thus, the pilot sets up these questions: If these results remain consistent in future iterations of this assessment, are the results satisfactory for these business students? Is an average evaluation of *good* by nine evaluators an acceptable standard of achievement? Should even one report be judged as having gross mistakes? The idea driving LeBow's assessment is improvement, and a key component of AACSB accreditation is that results should demonstrate development over time; ideally, then, scores would rise after LeBow implements specific pedagogical interventions in its curriculum, perhaps via faculty development in writing across the curriculum (WAC).

Of course, our method has limitations. All elements were weighed equally in calculating the overall assessment score. Perhaps some elements,

such as purpose or main point, should be weighed more heavily than others, such as document design or appearance. In fact, determining the weight of these elements could help further stimulate conversation and thinking about the use and evaluation of writing for LeBow. Also, the scores for ethics and gross mistakes may have improperly inflated the overall scores because the evaluators judged these 2-point elements as a 1.0 on the vast majority of the reports. But because we did not norm our evaluators, we were not surprised that normative answers did not spring from the data (e.g., What really constitutes an excellent element?).

This ongoing assessment may help us gain insight into the way two groups view student documents. Statistically significant evidence showed a difference between business and English evaluators in their evaluations of student performance. Abbott and Eubanks (2005) made a similar discovery when comparing academic technical writing instructors and practitioners, finding that although "academics and practitioners draw on standard sets of concepts and values that are likely at work whenever academics and professionals compose, edit, revise, or evaluate texts" (p. 207), the two groups still "also displayed distinct ways of evaluating sample texts" (p. 172). For us, this difference is especially interesting at the level of particular traits, or elements. Although the two groups agreed on the 2-point elements of ethics and gross mistakes, for each of the 4-point elements, the English evaluators judged the reports to be better than business evaluators did (see Figure 9). This noteworthy difference spurs questions for further study, including Can we get to the root of such a distinct difference? And does this difference have implications for students? We plan to use a similar philosophy of recruiting and training assessors in future assessments. If the results are consistent with the pilot in regard to this intriguing difference between the two groups, we would be well positioned to explore reasons why these differences exist.

The Future of the Assessment

As Yancey and Huot (1999) noted, "Much of the learning produced in the name of assessment is very subtle in nature" (p. 12). Our process provides a window through which to view student writing at LeBow. When this assessment is conducted more broadly and the results are analyzed, that view should be even clearer. By repeating this process annually, we will build a narrative of the progress and development of student

writing at LeBow. And when linked with the next step, the specific curricular interventions responding to the assessment findings, the process will be useful in the accreditation-driven efforts to document student performance—and certainly demonstrate the seriousness of purpose required by AACSB.

The collaboration between the LeBow faculty, the Department of English faculty, and a software company demonstrates how different entities with an interest in assessing and improving student writing can work synergistically. Opportunities for such collaborations might increase as accreditors push educational institutions to measure often-elusive outcomes, especially in areas such as writing and communication. Yancey and Huot (1999) stated that "assessment is increasingly collaborative and democratic" and that assessment need not be an expert-driven endeavor; instead it should be a "curricular task" that all stakeholders share in and learn from (p. 12). Prior, Hawisher, Gruber, and McLaughlin (1999) described this assessment philosophy:

> The desire to evangelize WAC and convert colleagues in other cultures is diminishing, and in its place is emerging an interest in understanding and learning from those in other disciplinary cultures and in engaging in cooperative action and dialogue with them to enrich the experiences of students and instructors in our universities. (p. 207)

In our experience, the collaborative nature of this project was a marked strength of what we did.

In this article, I have described a pilot process of conducting a large assessment of student writing. In addition, I have reviewed granular writing data that were designed to help a business school meet accreditation demands and that might lead to pedagogical and curricular change. This summary of the preliminary results of our pilot effort shows (a) how the LeBow College of Business will receive detailed feedback about writing traits from which it can align pedagogical objectives and (b) how two groups of evaluators, one from the English department and one from the business environment, differed significantly in their evaluation of the same reports. LeBow's efforts in devising and supporting this assessment may yield unique data that will help those involved with the education of business students to improve the writing instruction these students receive.

Appendix
The LeBow College of Business Writing Rubric

Gross mistakes (binary)	This reader did not find gross mistakes that would immediately cast suspicion on the effort/expertise of the document's creator.	This reader did find gross mistakes that made the reader suspicious or feel negative about the document.
Ethics (binary)	This reader thinks the document appears ethical, including its use of evidence.	This reader thinks some aspects of the document may be unethical; for example, the document uses inappropriate evidence (including plagiarism), recommends morally questionable practices, or uses deceitful data, style, or visuals.

	Excellent	Good	Fair	Poor
Purpose/main point	This reader thinks that the writer's purpose is clear. The document has a clear focus. Also, the idea presented/advanced/argued is both interesting and feasible.	This reader thinks the writer's purpose is clear for the most part.	The writer often loses focus of the main point of the document. The feasibility of the idea presented is questionable.	This reader has a difficult time determining why the writer has created this document. The main idea seems uninteresting and perhaps even unreasonable.
Audience	The writer has written for a clearly defined audience and, in this reader's opinion, has addressed that audience expertly. Also, it appears the writer has expertly followed the directions of the assignment/task.	The audience of the document is clear. This reader thinks the writer has done a good job addressing audience.	The document's treatment of audience is somewhat confusing. The writer does not seem to understand the audience of the document. Also, the writer does not appear to have clearly understood the assignment/task directions.	This reader thinks that the writer's treatment of audience appears unprofessional and/or it is not clear who is being addressed. Also, the writer has not followed the directions for the assignment/task.

(continued)

Appendix (continued)

	Excellent	Good	Fair	Poor
Organization	This reader thinks the report has a clear organizational logic. Transitions between ideas are handled well.	While the report is organized effectively, this reader thinks the document's organization could be refined/tightened a bit (headings, better transitions, etc.).	This reader thinks the document must be organized more effectively, as readers may be confused or misled.	This reader finds little coherent structure in this document. No clear rationale is apparent for why the document is set up the way it is. The document is confusing.
Evidence	This reader thinks the writer has made excellent use of research and sources, helping strengthen/build the document's main point with this material.	This reader thinks the writer made good use of research and sources. In a few places the document's main point could have been strengthened with additional evidence.	This reader thinks the document would be substantially strengthened with more/ better evidence, and/or the evidence presented is formatted in a sloppy, distracting manner.	The document is weak because of a lack of evidence and support, and/or the evidence used is formatted so poorly that it's difficult to tell what is cited.
Sentence style: Flow of writing	This reader thinks the clear, concise writing in this document made it easy (and perhaps even enjoyable) to read. The writer used solid sentence construction and strong word choices.	This reader thinks the writing in this document is good, but perhaps the writer could have written a bit more clearly and/or written more concisely.	This reader thinks some of the writing is awkward and clumsy, and/or the writer uses weak word choice or unsophisticated sentence structure.	This reader thinks that much of the writing in this document is awkward, repetitive, and/or wordy. The writing was not engaging.

(continued)

Appendix (continued)

	Excellent	Good	Fair	Poor
Correctness: Grammar and writing mechanics	This reader noticed few errors, if any. The document is clear, and the writer shows considerable mastery of the language.	This reader noticed some grammatical/mechanical errors, but those errors did not interfere with the reader's understanding of the document's purpose.	This reader noticed numerous grammatical/mechanical errors, and those errors interfered at times with the reader's understanding of the document's purpose and/or caused the reader to question the skill and expertise of the writer.	This reader noticed many grammatical/mechanical errors. The reader felt the number of errors made the document difficult to understand, and the reader questioned the document's credibility and the writer's skill because of these recurrent mistakes.
Document design/ appearance	This reader thinks the document uses design elements (white space, titles and subtitles, font size and style, etc.) expertly to create a professional-looking document that would satisfy the audience's expectations for that type of document.	This reader thinks the document is clean, but the appearance could be improved to aid in the document's clarity and/or organization.	This reader thinks the document has an amateurish look to it and/ or is in need of a more professional appearance. The audience may be confused by the design of the document.	This reader thinks the document appears sloppy and unprofessional and that sloppiness will certainly cause the audience to be confused.
Visuals (tables, charts, pictures, etc.)	This reader thinks the document utilizes visuals—tables, charts, pictures, etc.—in an expert way.	This reader thinks the writer makes good use of visuals. Perhaps there are additional opportunities for the use of such material or the material that is used could be improved somewhat.	The writer has missed opportunities to use visuals and/or has used visuals in a sloppy, ineffective way.	The writer needs visuals to help clarify the document's purpose, and/or the visuals used are sloppy, inaccurate, or presented in an unethical manner.

References

Abbott, C., & Eubanks, P. (2005). How academics and practitioners evaluate technical texts: A focus group study. *Journal of Business and Technical Communication*, *19*, 171-218.

Applebee, A., Langer, J., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal*, *40*, 685-730.

Elbow, P. (1996). Writing assessment in the 21st century: A utopian view. In L. Bloom, D. Daiker, & E. White (Eds.), *Composition in the twenty-first century: Crisis and change* (pp. 83-100). Carbondale: Southern Illinois University Press.

Haswell, R. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication*, *17*, 307-352.

Huot, B. (1996). The need for a theory of writing assessment. In L. Bloom, D. Daiker, & E. White (Eds.), *Composition in the twenty-first century: Crisis and change* (pp. 112-115). Carbondale: Southern Illinois University Press.

Huot, B. (1999). Beyond accountability: Reading with faculty as partners across the disciplines. In K. B. Yancey & B. Huot (Eds.), *Assessing writing across the curriculum: Diverse approaches and practices* (pp. 69-78). Greenwich, CT: Ablex.

Huot, B. (2002). *(Re)articulating writing assessment.* Logan: Utah University Press.

Langer, J. (2001). Beating the odds: Teaching middle and high school students to read and write well. *American Educational Research Journal*, *38*, 837-880.

McLeod, S. (2003). Celebrating diversity (in methodology). In L. Bloom, D. Daiker, & E. White (Eds.), *Composition studies in the new millennium: Rereading the past, rewriting the future* (pp. 151-154). Carbondale: Southern Illinois University Press.

Prior, P., Hawisher, G., Gruber, S., & MacLaughlin, N. (1999). Research and WAC evaluation: An in-progress reflection. In K. B. Yancey & B. Huot (Eds.), *Assessing writing across the curriculum: Diverse approaches and practices* (pp. 185-216). Greenwich, CT: Ablex.

Sommers, N. (2002). *The Harvard study of undergraduate writing*. Retrieved March 9, 2008, from http://www.fas.harvard.edu/~expos/index.cgi?section=study

Taylor, T. (2003). A methodology of our own. In L. Bloom, D. Daiker, & E. White (Eds.), *Composition studies in the new millennium: Rereading the past, rewriting the future* (pp. 142-150). Carbondale: Southern Illinois University Press.

Whithaus, C. (2005a). [Review of the book *Coming to terms: Theorizing writing assessment in composition studies*]. *Assessing Writing*, *10*, 214-218.

Whithaus, C. (2005b). *Teaching and evaluating writing in the age of computers and high-stakes testing.* Mahwah, NJ: Lawrence Erlbaum. Retrieved March 20, 2007, from http://site.ebrary.com/lib/drexel/Doc?id=10103804&ppg=121

Williamson, M. (1999). Pragmatism, positivism, and program evaluation. In K. B. Yancey & B. Huot (Eds.), *Assessing writing across the curriculum: Diverse approaches and practices* (pp. 251-262). Greenwich, CT: Ablex.

Xue, Y., & Meisels, S. J. (2004). Early literacy instruction and learning in kindergarten: Evidence from the early childhood longitudinal study—kindergarten class of 1998-99. *American Educational Research Journal*, *41*, 191-229.

Yancey, K. B., & Huot, B. (Eds.). (1999). Introduction: Assumptions about assessing WAC programs. Some axioms, some observations, some context. In K. B. Yancey & B. Huot (Eds.), *Assessing writing across the curriculum: Diverse approaches and practices* (pp. 7-14). Greenwich, CT: Ablex.

Zhao, J. J., & Alexander, M. W. (2004). The impact of business communication education on students' short- and long-term performances. *Business Communication Quarterly*, *67*, 24-40.

Scott Warnock is an assistant professor of English and director of the Freshman Writing Program at Drexel University. His research and teaching interests include uses of technology in writing instruction and assessment. He also helped develop Waypoint writing assessment software.