Big Data, Learning Analytics, and Social Assessment1

by Joe Moxley, University of South Florida

Abstract

This article explores the value of using social media and a community rubric to assess writing ability across genres, course sections, and classes. Since Fall 2011 through Spring 2013, approximately 70 instructors each semester in the first-year composition program at the University of South Florida have used one rubric to evaluate over 100,000 student essays. Between Fall 2012 and Spring 2013, students used the same rubric to conduct more than 20,000 peer reviews. The rubric was developed via a datagogical, crowdsourcing process (Moxley, 2008; Vieregge, Stedman, Mitchell, & Moxley, 2012). It was administrated via *My Reviewers*, a web-based software tool designed to facilitate document review, peer review, and writing program assessment. This report explores what we have learned by comparing rubric scores by project and semester on five measures (Focus, Organization, Evidence, Style, and Format) by project, section, semester, and course and by comparing independent evaluators' scores with classroom teachers' scores on two assignments for two semesters. Findings suggest use of the rubric across genres, sections, and courses facilitates a high level of inter-rater reliability among instructors; illustrates ways a curriculum affects student success; measures the level of difficulty of specific writing projects for student cohorts; and provides a measure of transfer. WPAs and instructors may close the assessment loop by consulting learning analytics that reveal real-time, big-data patterns, which facilitate evidence-based curriculum decisions. While not an absolute measure of student learning or ability, these methods enable tentative mapping of students' reasoning, research, and writing abilities.

Keywords: big data, writing assessment, social pedagogy, datagogies, transfer, curriculum standardization, peer production, communal agency

---

Introduction

The ways in which crowdsourcing processes and peer-production tools can be used to develop a standardized curriculum for first-year composition programs have been explored in recent qualitative research (Vieregge, Stedman, Mitchell, & Moxley, 2012), yet the degree to which crowdsourced assessment processes and social assessment tools impinge on teacher feedback, peer feedback, and assessment practices in university writing programs has received less attention. While in the past it was difficult for instructors to share comments on individual papers or compare grading patterns, database-driven, social-software tools and crowdsourced assessment practices now make it easy for Writing Program Administrators (WPAs) and instructors to share in the co-development of assessment rubrics; additionally instructors can share in-text and endnote comments, common comments, and rubric scores by student, section, and course.

The discipline of Writing Studies--broadly defined as graduate students, adjuncts, lecturers, tenure-earning faculty, and tenured faculty whose research and scholarship focuses on the study of writing K-college and beyond ["its production, its circulation, its uses, its role in the development of individuals and societies, and its learning by individuals, social collectives, and historically emergent cultures" (Bazerman, 2002, p. 32)]--is at pivotal moment in its assessment practices. In the United States, educators are under increasing pressure to develop valid measures to assess the development of writing and reasoning abilities. We face known limitations with our existing assessment procedures: Students' responses to essay prompts in timed situations may not reflect their abilities when it comes to revising and developing a project over time. Students' scores on a handful of papers, whether scored via an assignment-based rubric or a generic rubric, may not accurately reflect their current or future capabilities as writers. However, new technologies are becoming available that are changing the ecology of assessment. Rather than happening after the fact, assessment systems are becoming part of the teaching and learning process, part of the digital experience. Database tools can now track *all* student drafts, peer comments on drafts, peers' rubric scores, teachers' comments, and teachers' rubric scores. Hence, this study investigates the practice of using rubrics in a social media environment. In particular, it revisits the question regarding the value of using a generic rubric across genres, sections, and courses.

The Pros and Cons of Rubrics

On the one hand, since the late 1960s, assessment experts, teachers, and WPAs have praised rubrics for clarifying grading criteria and enabling multiple evaluators to reach unprecedented levels of inter-rater reliability (Broad, 2000). In general proponents of rubrics praise them for making evaluative criteria transparent and scoring more reliable. In his review of literature on rubrics, Bob Broad identifies Godshalk, Swineford, and Coffman (1966), Diederich (1974), and Cooper (1977) for being the first to pioneer rubrics as a way to improve score reliability among diverse readers, following Diederich, French, and Carlton's (1961) research that found even trained teachers rarely agreed about the quality of texts unless they used rubrics. Nowadays, rubrics invariably play starring roles in large-scale studies of writing (e.g., Arum & Roksa, 2011; U.S. Department of Education, 2012). Pagano, Bernhardt, Reynolds, Williams, & McCurrie (2008) found using a community rubric to conduct an inter-institutional analysis of student writing

between four universities helped instructors identify objectives and measure how student cohorts were doing in relation to one another. In summary, usage of rubrics has become ubiquitous across all levels of education.

On the other hand, some assessment experts and distinguished writing professors fault rubrics for oversimplifying or confounding the process of interpretation and response. Throughout his career, for instance, Peter Elbow has argued that assessment of writing is so subjective that teachers should be wary of assigning grades to specific texts and distrustful of grades or rankings they assign; rather than assigning a grade or using an analytic rubric to score specific criteria, Elbow prefers narrative responses to multiple texts in portfolios, particularly responses that are corroborated by multiple teachers/reviewers. For Elbow (1993), rubrics and training programs to teach readers to use rubrics misrepresent the diversity of readers' potential response styles:

> We can sometimes get agreement among readers from some subset, a particular community that has developed a strong set of common values, perhaps one English department or one writing program. But what is the value of such a rare agreement? It tells us nothing about how readers from other English departments or writing programs will judge--much less how readers from other domains will judge. (p. 189)

Likewise, Maja Wilson (2007) writes "rubrics meet the demands of objectivity by distancing teachers from their own perceptions in order to create agreement among readers--writing assessment's view from nowhere" (Rubrics and the View from Nowhere, para. 1).

Using a rubric to assess student work, particularly an early draft or the draft of a struggling student, may intrude on the teacher's ability to emphasize the feedback the student may need. For example, if a student has failed to understand the assignment or if a student clearly lacks an understanding of how to develop an argument for a critical reader, then it can be more useful to focus on providing feedback that addresses these specific deficiencies rather than providing more diffuse feedback. Clearly, the recommendation to identify one or two major patterns of error and to prioritize several global and local concerns can bump up against the practices of an analytical rubric that calls for commentary and evaluation on multiple criteria such as an assignment-based rubric might. Furthermore, students need a diversity of responses: they may grow bored with receiving a rubric response every time they write an essay if that is the only feedback they receive. As Elbow and Belanoff (2000) suggested in *Sharing and Responding*, sometimes students require no response other than the sound of their reading an essay out loud whereas other times they can benefit most from a face-to-face conference, a conceptual map of their organization, and so on.

Because different readers, discourse communities, genres, and media instantiate different assessment criteria, multiple rubrics are needed to assess student writing. As writing students face new writing challenges, as they are introduced to writing in their disciplines and workplace settings, they learn new genres of communicating, new methods of knowledge making, and new conventions for deploying evidence. Given this, the trend in Writing Studies has been away from using a single rubric to assess writing, research, and information literacy. In 1991, Richard Haswell rejected the use of rubrics for placement tests, arguing that natural development in writing occurs not across the board (as measured by holistic scales) but in sub-areas of writing competence. In a 2011 exchange on the WPA Listserv, Chris Anson concluded: "[The] [P]roblem with rubrics is their usual high level of generalization (which makes them worthless)." In a related scholarly analysis regarding the practice of using generic rubrics to assess the development of writing and reasoning abilities, Anson, Dannels, Flash, and Housley (2012) concluded, "generic, all-purpose criteria for evaluating writing and oral communication fail to reflect the linguistic, rhetorical, relational, and contextual characteristics of specific kinds of writing or speaking that we find in higher education." On the topic of generic rubrics, Bill Condon (2011) agreed with Anson on the same listserv discussion, suggesting "[generic rubrics] are not neutral, they are not local, and they cause far more problems than they solve--the greatest being, imho, their inherent reduction of the construct writing."

The disdain on the part of assessment experts regarding the practice of using a generic rubric to make global claims about writing development may be traced to past national research studies that have oversimplified the assessment of student reasoning and writing. For example, research such as Richard Arum and Josipa Roksa's *Academically Adrift* (2011) seems more harmful than productive. Using rubrics to assess a few paragraphs written by a small sample size (2300) of freshman and sophomore students writing outside of the scope of their graded, academic work, Arum and Roksa made broad claims regarding the efficacy of U.S. higher education: "[I]n terms of the general analytical competencies assessed, large numbers of U.S. college students can be accurately described as academically adrift" (2011, p. 121). Clearly, big rubrics can be used in deceptive ways--ways that are opaque, hyperbolic, and political.

Researchers in the area of knowledge transfer and threshold concepts have been particularly wary regarding the practice of using generic rubrics to make global claims about writing development. Bazerman (1988), Beaufort (2007), Carroll (2002), Ebest (2005), Nowacek (2011), Wardle (2009), Wardle & Roozen (2012) have explored ways that the ability to write well is not something that is mastered in one context and then simply carried over to another context. Clearly, arguments that make grand claims about student ability based on a handful of rubric scores need to be seriously challenged. Students' scores on one rubric are not necessarily predictive of how they will do when facing alternative genres.

Context

The First Year-Composition (FYC) Program at the University of South Florida serves approximately 4500 undergraduate students each year in two courses: ENC 1101 and ENC 1102. For 2011-2012, CCCC awarded the FYC Program the *Writing Program Certificate of Excellence*, a national award. The program's standardized curriculum, which is publicly available at http://fyc.usf.edu, is co-developed and under incessant revision by administrators and instructors in the program, who typically are graduate students in a doctoral rhetoric and composition, a Ph.D./MA program in literature, or an MFA program in creative writing. About 90% of our courses are taught by graduate students; the other 10% are taught by adjuncts who hold a minimum of an MA in English or an equivalent discipline and a record of successful college-level teaching.

The Development of the Community Rubric

As reported in a book-length, qualitative study, *Agency in the Age of Peer Production* (Vieregge et al., 2012), administrators and faculty at the University of South Florida have employed datagogical, crowdsourcing processes to develop a community rubric; associated resources for rubric criteria (videos on rubric terms, sample marked-up papers); and common comments for rubric criteria that can be embedded as hyperlinks on student papers. These common comments link out to an extended definition of the common comment, a video on the comment, two activities/exercises, and a link to a more extended essay related to the comment at *Writing Commons*, http://writingcommons.org, the open-education home for writers. Following a datagogical process, administrators and faculty have also developed a standardized curriculum for two courses, ENC 1101 and ENC 1102 (see http://fyc.usf.edu); and two original, textbooks (available in ebook and .pdf formats). Here, the term "datagogical" refers to

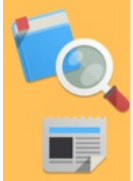
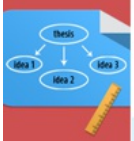
> what happens when "crowds" of teachers, students, and administrators use social software to develop pedagogical communities that value and are fueled by the "wisdom of crowds"--the surprising ability of crowds of people to develop pedagogies that are wiser and more engaging than those developed by individuals, even disciplinary experts. Rather than being theorized by experts, vetted by the peer-review process, and published after a long wait, datagogies are pedagogies that are subject to immediate revision, collaboration, and even deletion. Via the datagogy, users--other teachers and students--can develop pedagogical practices in real time. Datagogies can challenge traditional assumptions about authorship, authority, collaboration, and power. Teaching, learning, and writing can become more dialogical as opposed to presentational. Knowledge can be conditional, subject to the next edit. Datagogies have the potential to dramatically alter collaboration, creativity, and community. (Moxley, 2008, pp. 182-183).

After participating in ongoing discussions about our curriculum and shared evaluative criteria, FYC staff and teachers came to conceive of our effort as a crowdsourced, *community rubric* as opposed to an *institutional rubric* or *generic rubric*. For us the term *generic* or *institutional rubric* was associated with too many negative connotations, conveying the sense of standards delivered down from some legislative or professional body. In contrast, our rubric and supporting educational resources were developed in response to listserv discussions, face-to-face arguments, surveys, interviews, and informal teacher feedback.

During the fall of 2011 through spring of 2013, the five major sections of our rubric remained the same: *Focus* (Basics and Critical Thinking), *Evidence* (Critical Thinking), *Organization* (Basics and Critical Thinking), *Style* (Basics and Critical Thinking), and *Format* (Basics). As indicated by Figure 1, the term *Basics* refers to adherence to Standard English or basic conventions, such as correct use of verbs or formatting conventions. In contrast, *Critical Thinking* refers to more rhetorical, global issues such as organizational flow across paragraphs.

While our evaluative criteria can be traced to Paul Diederich's (1974) early work on analytic rubrics, our instructors have expressed pride and a sense of ownership about our rubric and related resources because they contributed significantly to their development. While our approach--a standardized curriculum and rubric--may limit agency at the individual level, it can also create a new form of communal agency, as analyzed in *Agency in the Age of Peer Production* (Vieregge et al., 2012). Likewise, renaming our rubric a *community rubric* as opposed to an institutional rubric doesn't suddenly overcome the weaknesses of rubrics identified by assessment leaders, although communal approaches such as ours may be preferable to traditional approaches--that is, downloading a rubric from a professional organization such as the Council of Writing Programs or Association of American Colleges and Universities.

Figure 1. Community rubric.

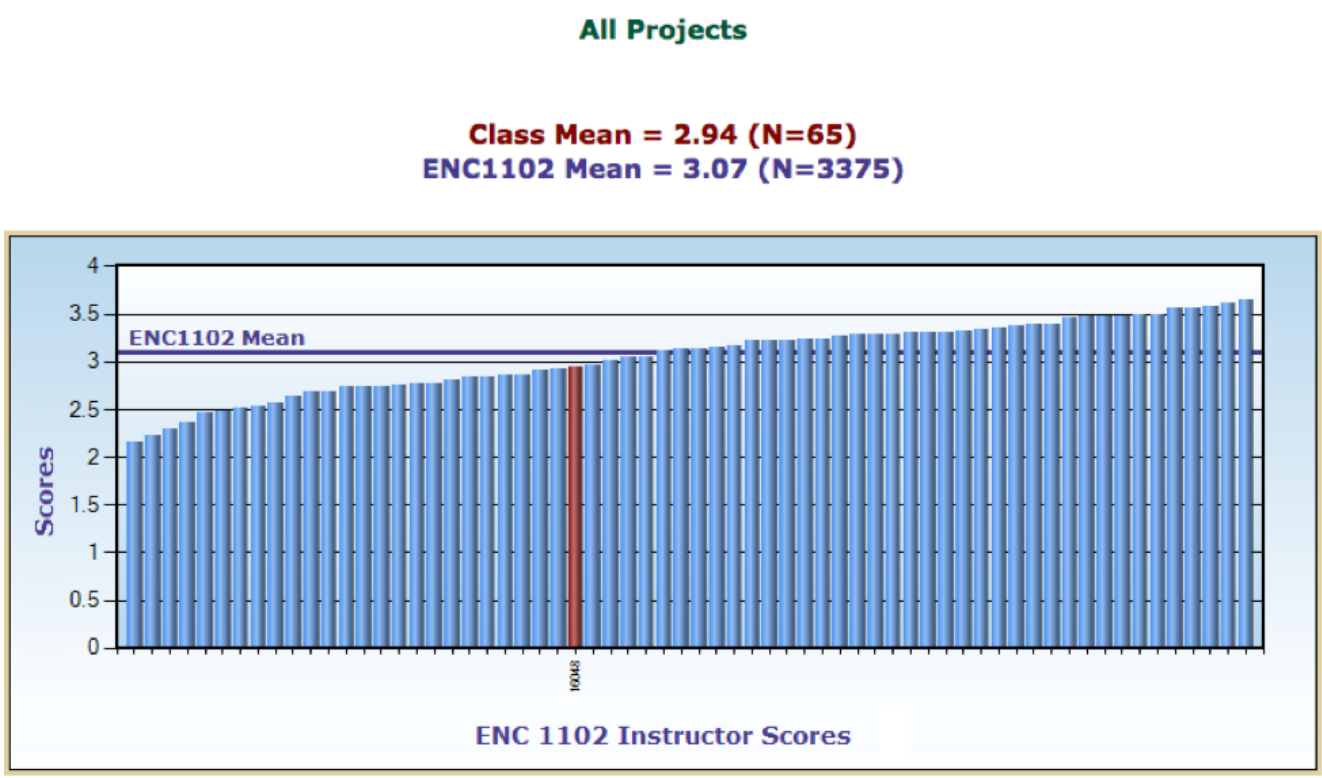
Criteria	Level	Emerging 0	1	Developing 2	3	Mastering 4
Focus 	***Basics***	Does not meet assignment requirements		Partially meets assignment requirements		Meets assignment requirements
	Critical Thinking	Absent or weak thesis; ideas are underdeveloped, vague or unrelated to thesis; poor analysis of ideas relevant to thesis		Predictable or unoriginal thesis; ideas are partially developed and related to thesis; inconsistent analysis of subject relevant to thesis		Insightful/intriguing thesis; ideas are convincing and compelling; cogent analysis of subject relevant to thesis
Evidence 	***Critical Thinking***	Sources and supporting details lack credibility; poor synthesis of primary and secondary sources/evidence relevant to thesis; poor synthesis of visuals/personal experience/anecdotes relevant to thesis; rarely distinguishes between writer's ideas and source's ideas		Fair selection of credible sources and supporting details; unclear relationship between thesis and primary and secondary sources/evidence; ineffective synthesis of sources/evidence relevant to thesis; occasionally effective synthesis of visuals/personal experience/anecdotes relevant to thesis; inconsistently distinguishes between writer's ideas and source's ideas		Credible and useful sources and supporting details; cogent synthesis of primary and secondary sources/evidence relevant to thesis; clever synthesis of visuals/personal experience/anecdotes relevant to thesis; distinguishes between writer's ideas and source's ideas
Organization 	***Basics***	Confusing opening; absent, inconsistent, or non-relevant topic sentences; few transitions and absent or unsatisfying conclusion		Uninteresting or somewhat trite introduction, inconsistent use of topics sentences, segues, transitions, and mediocre conclusion		Engaging introduction, relevant topic sentences, good segues, appropriate transitions, and compelling conclusion
	Critical Thinking	Illogical progression of supporting points; lacks cohesiveness		Supporting points follow somewhat logical progression; occasional wandering of ideas; some interruption of cohesiveness		Logical progression of supporting points; very cohesive
Style 	***Basics***	Frequent grammar, punctuation errors; inconsistent point of view		Some grammar/punctuation errors occur in some places; somewhat consistent point of view		Correct grammar and punctuation; consistent point of view
	Critical Thinking	Significant problems with syntax, diction, word choice, and vocabulary		Occasional problems with syntax, diction, word choice, and vocabulary		Rhetorically-sound syntax, diction, word choice, and vocabulary; effective use of figurative language
Format 	***Basics***	Little compliance with accepted documentation style (i.e., MLA, APA) for paper formatting, in-text citations, annotated bibliographies, and works cited; minimal attention to document design		Inconsistent compliance with accepted documentation style (i.e., MLA, APA) for paper formatting, in-text citations, annotated bibliographies, and works cited; some attention to document design		Consistent compliance with accepted documentation style (i.e., MLA, APA) for paper formatting, in-text citations, annotated bibliographies, and works cited; strong attention to document design

The Development of My Reviewers

My Reviewers, a document workflow, peer review, e-portfolio tool, was developed with support from the University of South Florida.[2] Teachers and students use *My Reviewers* to assess student work with a rubric, to make comments on students' papers; and to choose from a library of comments that can be embedded on top of student papers. Teachers can also use *My Reviewers* to establish and monitor peer-review teams; and they may consult onboard learning analytics that aggregate results (endnotes, grades, comments) by section and class. These learning analytics provide *social-software features* such as the ability for *administrators* to view instructors' commenting pacing as well as instructors' scores and comments; for *instructors* to view students' peer reviews and compare aggregated-peer-review data (in-text comments, end-note comments, and rubric scores); for *instructors* to view across their

sections or across other sections of the same course; and for *students* to view peer-reviews (including interface views that publish alternative reviews in contrast with one another) and to rate reviewers' feedback.[3] These social tools create a high-level of accountability and permissions-based transparency for user roles (*administrators*, *mentors*, *instructors*, and *students*). For example, as illustrated by Figure 2 below, instructors can access the analytics to see how their grades compare with the grades other instructors are providing to students in the program on the same projects. Or WPAs can refer to the document upload or peer review analytics to ensure teachers are keeping pace with the standardized curriculum.

Figure 2. Illustration of an instructor's view of his/her score in relation to other instructors' scores for all projects. This example shows the view of the instructor for ENC 1102 section 16048, reporting his or her mean score on 65 projects was 2.94 while the mean score on 3375 projects by other teachers in the program that semester was 3.07.



Traditional Assessment Methods vs. Big Data Methods

Until the recent development of database assessment tools like *My Reviewers*, researchers had difficulty working with large datasets. As a result, researchers have used relatively small sample sizes, typically comparing rubric scores on one set of papers against another set regardless of the genre or exigency of the writing task. For example, while Arum and Roksa's research methods have largely been discredited (Astin, 2011; Glenn, 2011; Haswell, 2012; Hosch, 2010), perhaps the most damning critique is tied to the evidence they provide for their claim that undergraduates aren't learning: the Collegiate Learning Assessment (CLA), the test on which they base their argument, employs open-ended questions to measure reasoning and writing skills, which may not reflect what students learn about disciplinary ways of making and assessing knowledge. Plus, we can't be sure how seriously the 2,300 students in the study took these questions, especially given these tests were not an integral part of the students' required coursework.

Even the U.S. Department of Education's (2011) "The Nation's Report Card: Writing 2011," a study conducted by the National Assessment of Educational Progress, which used rubrics to score a representative random sampling of 22,000 American high school and middle school students who wrote two 30-minute essays, seems severely flawed: These essays were written outside of students' graded, day-to-day work, and they represent a distorted view of the composing process; few of us write important academic documents in 30 minutes. Locally, my university's efforts, conducted by the Office of Institutional Effectiveness, which asked independent raters to use our community rubric to score the final essays written by 5% of our population in 1102 with the first papers these same students wrote in 1101, illustrate some of the same problems that Broad (2003), Anson et al. (2012), and Condon (2012) identified: Conclusions are being made about the development of students' reasoning and writing abilities based on students' performance on two papers that address remarkably different genres. Comparing students' scores on a rhetorical analysis project with a Rogerian argument is like comparing apples to papayas.

In contrast, big-data assessment methods provide an important alternative to traditional assessment practices, such as those employed by Arum and Roksa or the National Assessment of Educational Progress. Powered by databases that archive *all*

instructors' and peer reviewers' rubric scores and comments written on their papers or endnotes and inserted common comments, big-data methods capture a great deal more information, thereby avoiding the sorts of simplistic conclusions that have undermined traditional assessment methods. To summarize, big data research methods in partnership with digital assessment tools archive all student work conducted during the day-to-day work of their courses.

Method

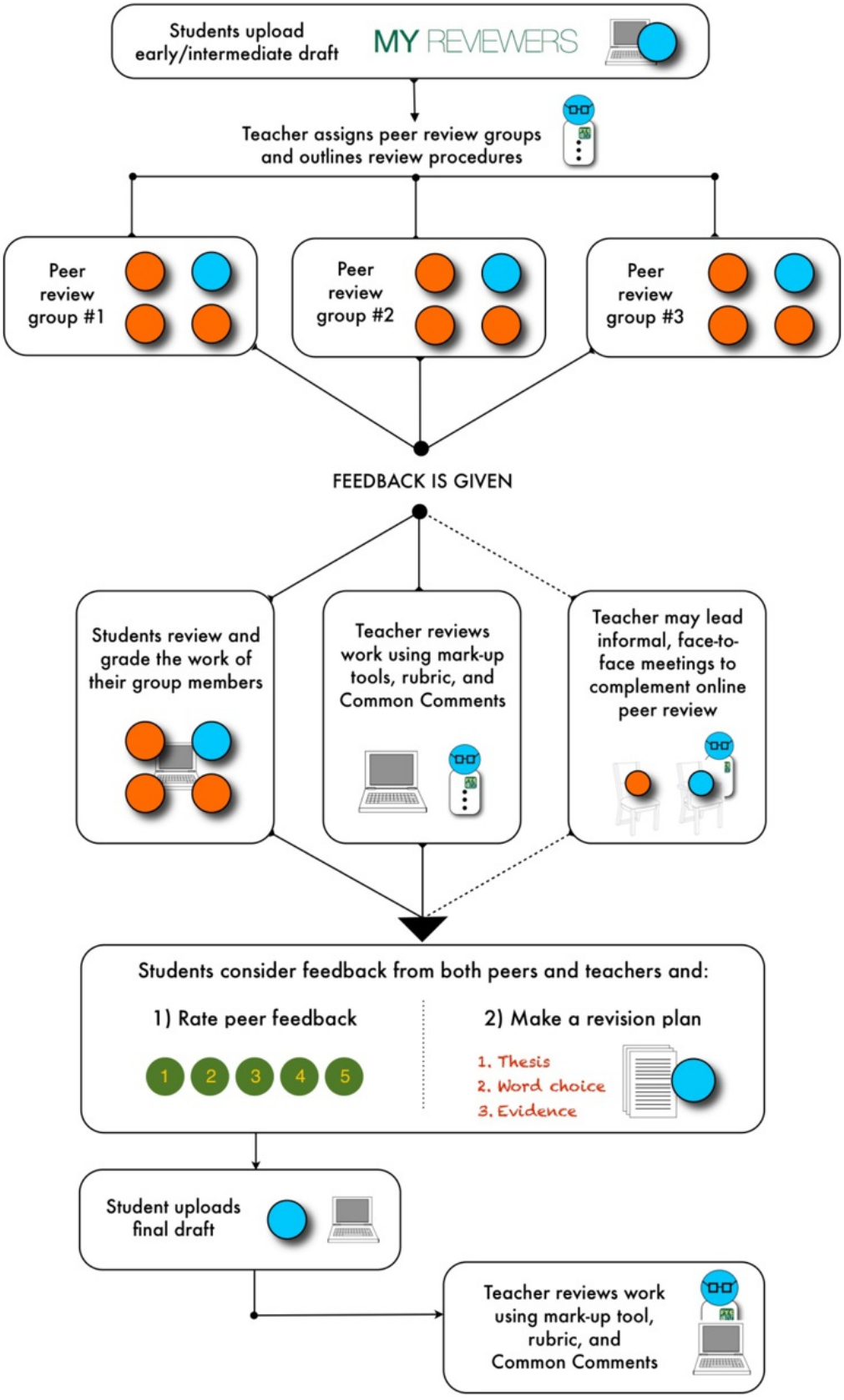
Two methodological approaches were followed: First, as reported in "Aggregated Assessment and 'Objectivity 2.0'" (Moxley, 2012), to determine inter-rater reliability among instructors, in the summer of 2011, the Office of Institutional Effectiveness, an independent office at our university, hired 10 English Instructors to score randomly chosen essays for Project 2 in ENC 1101 with these same students' essays for project 2 in ENC 1102. In particular, 249 essays written for Project 2 in ENC 1101 and 249 essays written by the same students for Project 2 in ENC 1102 were read by both the classroom teachers (as part of routine teaching work) and by independent scorers (who scored without knowing the classroom instructors' scores). This sample size was randomly chosen, constituting 5% of our total population that year (based on unique student count rather than total student count).

Second, instructors scored all of their students' intermediate drafts and final drafts with the community rubric. Teachers and administrators were then able to access the Learning Analytics to view various big data trends. For example, *My Reviewers* sums and averages instructors' scores on the rubric criteria (Focus, Organization, Evidence, Style, and Format) for all submitted intermediate and final drafts by project for a particular semester.

Participants

Since Fall 2011 through Summer 2013 instructors in the first-year composition at the University of South Florida used one rubric to evaluate over 100,000 student essays. Between Fall 2012 and Summer 2013, students used the same rubric to conduct more than 20,000 peer reviews. Usage of the rubric varied from semester to semester. During the first academic year of development (Fall 2012 and Spring 2013), students were required to upload final drafts. Since Spring 2010, students were required to upload both intermediate and final drafts. Peer review features were added as an option in Spring 2011. Since Spring 2012, students have been required to conduct a minimum of six peer reviews--two for each major paper. Instructors determine group sizes, from two to five students per group. Typically intermediate drafts and peer reviews were discussed in face-to-face meetings as well as small group settings.

Figure 3. Illustration of workflow for *My Reviewers*.



Data Collection: The Program Corpus vs. the Research Corpus

According to our agreement with the USF Institutional Review Board, the data (e.g., all rubric scores, in-text comments, and endnotes) provided to researchers do not include student and instructor identification numbers or names. Instead, *My Reviewers* substitutes an alternative, randomly generated ID to organize records for researchers. So long as researchers cannot associate student identities with data, USF, in consultation with Western IRB, ruled this research was exempt from IRB. Even so, as an ethical matter, when students upload documents they are prompted to opt in or opt out of having their work added to a corpus that can be data mined for research purposes. Historically, less than two percent opt out. Thankfully, the learning analytics tools at *My Reviewers* provide administrators access to all aggregated data for program assessment as well as reports containing data from

students who approved adding their texts to the *My Reviewers* corpus for research purposes. The latter corpus is used in this study.

Study Limitations

The narrow focus in this study on ways WPAs can use real-time analytics to measure student reasoning and writing provides an incomplete mapping of our assessment ecology, similar to characterizing a singular athletic event at a sports game, such as a football touchdown or soccer goal, as the meme for the entire game. After all, this focus on the efficacy of using a rubric across genres, course sections, and courses limits discussion of other chapters of our story. For example, this research does not explore privacy issues surrounding big-data assessment approaches or research on what we can learn from hundreds of thousands of teacher and student comments on papers and endnotes. This study doesn't analyze what we can learn about peer review by comparing students' and teachers' comments and endnotes and rubric scores on students' peer reviews. This study doesn't explore how teachers and students feel about the pancopticon-esque nature of our assessment ecology nor does the study investigate what inspires instructors to contribute (or not contribute) to our ongoing efforts to develop community-based pedagogical resources. The study doesn't explore how students use the archive of teacher and peer feedback to identify recurring weaknesses in students' reasoning, research, or writing nor does it track feedback on students' papers or rubric scores from a longitudinal perspective. Moreover, this study doesn't evaluate instructors' use of onboard social tools, such as the ability to compare their grades against other instructors' grades in the program.

Study limitations extend beyond my narrow focus on learning analytics pertaining to students' final draft scores. After all, while more than 120,000 essays have been scored by instructors or students without direct input on my part, my interpretations regarding what these rubric scores mean are surely affected by my positionality as the WPA for the program and chief architect of *My Reviewers*. While I clearly cannot step outside my positionality, I do endeavor to be as objective as possible by providing a detailed description of the curriculum instructors employed when evaluating each set of papers, by considering alternative, contrary interpretations, and by acknowledging the murky areas--the spaces where I cannot fully explain the results.

Findings

Comparison of rubric scores by project sequence, section, and course and comparison between independent evaluators' scores and classroom teachers' scores reveals advantages to big data, corpus-based assessment methods: Surprisingly, use of the community rubric across genres, course sections, and courses--at least for the two semesters analyzed below--seems to have enabled instructors in the writing program to grade students' work in equivalent ways. Furthermore, this practice provides a baseline measure of a particular group's reasoning and writing abilities. WPAs can make evidence-based curriculum changes in response to real-time assessment results and then compare other cohorts' baseline performances. The finding, discussed below, that particular cohorts' aggregated scores rise for personal narratives and reflection essays and fall for research-based essays offers a snapshot of the complexities of knowledge transfer and the development of reasoning and writing abilities.

"Objectivity 2.0"

Obviously, a tool isn't going to change human nature, suddenly transforming what is an inevitably subjective process into an objective one. That said, as Yochai Benkler (2006) has argued, when a tool makes a process easier, it's more likely to be accomplished. The social features of digital assessment tools that remediate traditional assessment practices--for example, the ability of administrators to view instructors' scores and comments along with the ability for instructors to easily view all student peer reviews--combined with the use of the crowdbased community rubric and rubric resources (videos of rubric terms and peer-review tips, sample marked up papers, and so on) may make it more likely instructors could reach agreement with one another.

In the summer of 2011, when the Office of Institutional Effectiveness compared the scores assigned by ten independent raters with students' classroom teachers' scores, they found no significant differences on 7 of 8 rubric measures (Moxley, 2012). The only real difference between the students' classroom instructors and the independent evaluators was that the *Style: Basics* criterion was graded more harshly by the students' classroom instructors than by the independent evaluators. While the high level of inter-rater reliability among the independent scorers (.93) during their preliminary training from our institution's Office of Institutional Effectiveness may not be terribly surprising, the finding that the independent evaluators' scores were relatively equivalent to the students' classroom teachers' scores was quite remarkable and unexpected.

Without a follow up study or additional context-based research, I cannot fully account for the judgment processes of instructors and independent raters. I don't know enough about how our teachers or independent assessors navigate the gray area between the generalizations embedded in our community rubric and the particular outcomes of our projects. In 2010, I know, for example, that the *Organization* for Project 2 in 1101 (a historiography project) differed from Project 2 in 1102 (a Rogerian argument) but I don't know how teachers' understanding of the various project outcomes weighs in their decision-making process when using the community rubric. However, the strong inter-rater reliability of our instructors and independent evaluators during Fall 2010 and Fall 2011 for Project 2 suggests a community rubric tied to a document workflow tool such as *My Reviewers* can enable diverse teachers to

reach what we might call "Objectivity 2.0"--a form of aggregated objectivity.

While preliminary and speculative, these results provide empirical support for Gerald Graff's contention that students benefit when teachers open their classroom doors and speak with other teachers about desired shared outcomes. As Graff argues in *Clueless in Academe* (2003) and "Why Assessment?" (2010), when faculty focus solely on what goes on in their classrooms, when they shun discussions with colleagues regarding shared criteria for evaluating writing, they may create "trickledown obfuscation" which may impede students from joining and understanding academic communities (2010, p. 157).

Transfer and Evidence-Based Curriculum Revisions

Beyond enhancing "Objectivity 2.0" among our instructors, big-data assessment methods provide a lens to evaluate the efficacy of a curriculum. Within the confines of a semester, WPAs can analyze students' aggregated rubric scores to develop a baseline measure of how a particular cohort of students does on a particular assignment sequence. When aggregated scores decline or are sporadic this may indicate students are having difficulty transferring what they have learned from previous assignments to new assignments. Furthermore, after making curriculum changes, WPAs can compare aggregated rubric scores across student cohorts. Interestingly, this process helps us see how grades are determined as much by the curriculum as by student ability, thereby offering a counterintuitive insight into student success.

For example, in the fall of 2011, my colleagues and I were surprised when the aggregated rubric results for students in ENC 1101 failed to improve: The 1065 students who completed Project 1, a personal narrative, scored 3.1 overall. Subsequently, for the 1033 students who completed Project 2, a literature review, the overall score dipped to 2.8. Then the scores of the 1016 students who completed Project 3, a second personal reflection, improved somewhat: 3.3. But the 1102 students who completed Project 4, a thesis-driven essay, returned to where they began the semester: 3.1. At the time, our sense was that the semester had been worthwhile, yet we were disappointed that students' scores didn't really improve during the semester. While we didn't necessarily define success as ever higher scores from project to project, were still troubled by the lack of progress, particularly given that we typically witnessed considerable progress in ENC 1102, our second-semester composition course.

To examine the score differences among projects, repeated ANOVA tests were conducted for each criterion. Table 1 summarizes the rubric scores by criteria as well as the results of repeated ANOVA tests. The test results demonstrated a significant effect of projects on student scores for every writing skill. The following tests for pair-wise comparisons of the ANOVA tests showed the statistically significant differences (at α=.05) among scores for all 4 projects.

Table 1. *Descriptive Statistics and ANOVA Analysis Results by Criteria in ENC 1101 sections during the Fall 2011 Semester*

Criteria	Skill	Project 1 n = 1065	Project 2 n = 1033	Project 3 n = 1016	Project 4 n = 1102	F-value
Focus	Basic	3.74 (0.62)	3.36 (0.99)	3.51 (0.88)	3.51 (0.94)	31.89***
	Critical thinking	2.97 (0.84)	2.87 (0.96)	3.08 (0.85)	2.98 (0.92)	9.82***
Evidence	Critical thinking	3.06 (0.85)	2.94 (1.01)	3.04 (1.01)	3.03 (0.99)	3.07***
Organization	Basic	2.89 (0.85)	2.92 (0.88)	3.05 (0.83)	3.04 (0.85)	15.94***
	Critical thinking	2.94 (0.83)	2.98 (0.91)	3.1 (0.82)	3.07 (0.88)	13.92***
Style	Basic	2.84 (0.88)	2.87 (0.88)	3.04 (0.84)	2.94 (0.85)	16.26***
	Critical thinking	2.72 (0.86)	2.74 (0.88)	2.86 (0.82)	2.8 (0.84)	9.05***
Overall Rubric		3.04 (0.6)	2.95 (0.75)	3.09 (0.68)	3.04 (0.72)	12.84***

After some discussion, my colleagues and I decided we should give less classroom time in Fall 2012 to teaching the personal narrative and more time to teaching academic discourse, particularly historiography and rhetorical analysis. Rather than four assignments that oscillated between personal and academic discourse, we settled on three more traditional academic projects: a bibliographic essay, a thesis-driven essay, and a remediation essay. Subsequently, at the conclusion of the fall 2012 semester we analyzed our big data trends and compared them with the previous year's trends.

A series of repeated ANOVA test was conducted to compare project scores by each writing skill and overall rubric of students in

ENC 1101 courses during the fall 2012 semester. Table 2 shows the test outcomes as well as the standard deviation of scores for all three projects. The results indicated that scores of three projects illustrated statistically significant difference at 99% confidence level for all writing skills except for *Style: Basics* skill, which displays a statistically significant difference at 90% confidence level. The following tests for pair-wise comparisons of these ANOVA tests were also statistically significant at alpha level of .05.

Table 2. *Comparison of Projects' Scores By Skill and Total Weighted Score for Students in ENC 1101 sections during the Fall 2012 Semester*

Criteria	Skill	Project 1 n = 1061	Project 2 n = 1017	Project 3 N = 993	F-Value
Focus	Basic	3.17	3.54	3.41	48.66***
		(1.11)	(0.8)	(0.97)	
	Critical thinking	2.79	2.96	3.11	40.28***
		(1)	(0.9)	(0.93)	
Evidence	Critical thinking	2.84	2.93	3.03	12.19***
		(1.06)	(0.96)	(0.99)	
Organization	Basic	2.92	3.1	3.12	22.86***
		(0.92)	(0.83)	(0.83)	
	Critical thinking	2.98	3.25	3.24	47.57***
		(0.94)	(0.83)	(0.83)	
Style	Basic	2.91	2.98	3.01	2.43*
		(0.92)	(0.82)	(0.81)	
	Critical thinking	2.8	3.03	3.03	37.69***
		(0.9)	(0.83)	(0.82)	
Overall Rubric		2.9	3.12	3.13	64.11***
		(0.77)	(0.64)	(0.67)	

Note. * = p<0.1, *** = p<.0001. Standard deviations appear in parentheses below the means. Means are significant different at p<.05 based on Contract Post-hoc paired comparisons

This process reveals ways big data assessment methods can inform evidence-based curriculum revisions; however, it doesn't resolve some of the larger issues: How concerned should WPAs be when aggregated project scores stay flat or decline across the semester? When students' scores decline, can this still be beneficial to students' development? Are WPAs helping students or disserving them by scaffolding a curriculum that is more likely to result in higher overall scores?

Evidence of Student Reasoning, Information Literacy, and Writing

In contrast to Arum and Roksa's argument that U.S. students don't learn over two years of undergraduate coursework, I suspect most writing teachers routinely witness student improvement as our students work from early drafts, through intermediate drafts, to final drafts. Though we can feel our students develop, though we can see it in their faces when we work with them, how can we qualify and quantify that development?

At what point can we really assess students' reasoning and writing abilities? Should our measure of growth begin with students' first drafts because they represent what students can do by themselves? Are final drafts, which have developed in response to help from teachers, peers, (and sometimes writing center tutors), an accurate representation of student ability? To determine whether students are improving as writers, should we compare teachers' scores on students' intermediate drafts and final drafts? Or, since even professional writers' intermediate drafts can be extremely rough, are intermediate drafts too soft of a ground to measure?

The ocean of data provided by digital assessment tools affords researchers new opportunities for measuring the development of critical thinking and writing abilities. As Haswell (2000) has argued, we can probably develop our best portrait of writing development by employing multiple measures. As we develop a global corpus of student work and teacher commentary or even as we look at corpi developed at our home institutions, we can explore a range of data collection and analysis tools. As a first step in a big data approach to measuring development, I thought it made most sense to initially compare students' aggregated final draft scores for one semester. That said, I admit one major problem with measuring growth based on final drafts is that students may have received loads of help to reach that final stage--from one-on-one conferences with their instructors, access to online via tools like Smart Thinking, and visits to our university's writing center.

Table 3 summarizes the standard deviations of project scores, as well as the results of the repeated measures ANOVA tests for seven writing skills, and the overall rubric score of ENC 1102 students in Fall 2012. The results show that the scores of these projects were statistically significant different at 99% confidence level for all reasoning and writing skills. The contract post-hoc tests of these ANOVA tests also reported statistically significant at the 95% confidence level. This is a statistical evidence for the improvement of the ENC 1102 students in reasoning and writing.

Table 3. *Comparison of Projects' Scores by Skill And Total Weighted Score for Students in ENC 1102 Courses during the Fall 2012 Semester*

Criteria	Skill	Project 1 N = 893	Project 2 N = 879	Project 3 N = 879	F-Value
Focus	Basic	3.24	3.3	3.46	13.80***
		(0.95)	(0.93)	(0.84)	
	Critical thinking	2.81	2.94	3.14	33.55***
		(0.97)	(0.92)	(0.89)	
Evidence	Critical thinking	2.64	2.89	2.98	29.31***
		(1.02)	(0.99)	(1.03)	
Organization	Basic	2.94	3.06	3.24	34.45***
		(0.93)	(0.89)	(0.82)	
	Critical thinking	2.86	3.07	3.27	59.96***
		(0.95)	(0.89)	(0.85)	
Style	Basic	2.8	2.93	3.06	25.41***
		(0.94)	(0.91)	(0.86)	
	Critical thinking	2.79	2.87	3.05	34.08***
		(0.92)	(0.92)	(0.86)	
Overall Rubric		2.84	2.99	3.15	66.39***
		(0.73)	(0.72)	(0.70)	

Note. *** = p<.0001. Standard deviations appear in parentheses below the means. Means are significant different at p<.05 based on Contract Post-hoc paired comparisons

To add depth to the statistical finding of improvement and to bring dimensionality to the overall portrait of development, I believe we need to employ multiple measures, including case studies of individual writers, linguistic analysis, and discourse analysis. We need additional research to understand how our teachers juggle the generalizations instantiated in our common rubric with the more unique requirements of diverse genres. Despite these important limitations, using real-time learning analytics seems more valid to me than one-shot assessment tests like the CLA or FCAT test in Florida. After all, in contrast to a couple of paragraphs written for the CLA or a brief essay written for the SAT, FCAT, or AP English, big data assessment methods are more likely to accurately measure student reasoning and writing because the work being assessed is conducted in the context of real, graded student work. Many teachers follow what we might call the *Writing Studies Assessment Manifesto: students care about an activity when they are graded on that activity (or in the case of portfolio-based approaches, know they will eventually be graded on their final products), particularly as they grow older and more cynical about meaningless academic tests*.

Summary

Just as writing assessment moved first from objective tests, to holistic scores of student papers based on rubrics, to portfolio assessments, to e-portfolios, and then large-scale writing program reviews--as noted by Kathleen Yancey (1999)--now the turn to big data assessment is inevitable. As Lang and Baehr (2012) mention in their essay on the topic,

> [w]riting program administrators, faced with increasing demands for accountability and assessment, as well as widely varying student populations, need to have ways of understanding the interactions of students, faculty, and administrators in their present program, both in the short term and longitudinally. (p. 173).

This study identifies substantive benefits to employing a single unit of measure, a community rubric, to assess intermediate and final writing projects in two composition courses. Consistent, repetitious use of a common rubric across genres, sections, course sections, and courses enhances inter-rater reliability among our teachers. While these findings do not bring clarity as to how reviewers use a criterion like "focus" to evaluate a rhetorical analysis differently from a literature review, remediation paper, historiographical analysis, or social action paper, these findings support Graff's argument that teachers across disciplines, particularly professors of general education courses, have much to gain by dialoging with one another regarding shared

conventions.

Ideally, used in conjunction with vertical approaches to writing program administration, faculty will extend these big-data assessment methods beyond a first-year composition program to include an entire general education program. As students work their way through courses, they could tag documents to be included in their final e-portfolio and then be asked in a capstone course to reflect on what they have learned about writing and research processes. Rather than assessing development on meaningless paragraphs written for tests like the CLA or on one or two sets of papers, assessment could be better integrated into teaching and learning processes.

Creating a culture of assessment in the FYC Program has been beneficial at USF. Real-time assessment information enables us to follow up big data patterns with evidence-based curriculum changes. Ongoing discussions about the community rubric and the development of resources for our students have kept us focused on our students' needs as writers, researchers, and critical thinkers. Now we have a database of approximately 120,000 documents, which is attracting global interest on the part of computational linguists and Writing Studies scholars. By following a do-it-yourself model for developing textbooks and assessment tools, writing programs can generate ongoing funds to pay for course-load releases for graduate students who can then work on curriculum development, e-texts, videos, podcasts, quizzes, and exercises. A university that uses learning analytics throughout the undergraduate curriculum can better prepare students for careers as writers, researchers, and citizens.

When NASA launched the Hubble space telescope into the atmosphere in 1990, scientists' initial excitement about seeing deeper into the solar system than had ever been possible was undercut by the realization that there was a major flaw in the design of the telescope. Thankfully, NASA was able to repair the telescope, thereby enabling scientists to rewrite astronomy and physics. Whereas our use of a common rubric within a document workflow, e-portfolio tool like *My Reviewers* is clearly not as sophisticated as the Hubble, it too represents a lens, a new way of mapping the subjectivities of interpretation and assessment--even if our lens is still a little out of focus, a little murky. As developers work on digital assessment tools and as researchers and higher education institutions pool students' texts, into a worldwide corpus, we will be able to see deeper into how individuals and communities assess texts and how students develop as writers, thinkers, researchers, and citizens. Learning analytics, tools that interpret this ocean of deep data, will help us better understand when and how to comment on student writing. Corpus-based research projects will help us address new research projects, from examining the efficacy of particular comments, evaluating and adjusting curricular approaches, to developing intelligent agents that respond in real time to students and instructors based on past performances and even suggesting exercises for students and instructors. Like other professions that are being remediated by social, networked knowledge-making practices, faculty members' assessment practices will become more social, transparent, and better coordinated.

Notes

[1] This research would never have been possible without the creativity and professional dedication of my colleagues at the University of South Florida. Most importantly, I thank Terry Beavers, the chief developer of My Reviewers; Dianne Donnelly, the associate director of the writing program; Ellie Bieze, Megan McIntyre, Jessica McKee, Erin Trauth, Taylor Mitchell, the mentoring coordinators over the past seven years; and Daniel Richards and Kyle Stedman, the community managers for FYC. My thanks to Gerald Graff, Norbert Elliot, Chris Anson, Diane Kelly-Riley, Peggy O'Neil, and Ann Damiano for their insightful comments on drafts of this research study. Finally, I thank Diep Nguyen for compiling the statistical analysis employed in this study.

[2] This research would never have been possible without the creativity and professional dedication of my colleagues at the University of South Florida. Most importantly, I thank Terry Beavers, the chief developer of My Reviewers; Dianne Donnelly, the associate director of the writing program; Ellie Bieze, Megan McIntyre, Jessica McKee, Erin Trauth, Taylor Mitchell, the mentoring coordinators over the past seven years; and Daniel Richards and Kyle Stedman, the community managers for FYC. My thanks to Gerald Graff, Norbert Elliot, Chris Anson, Diane Kelly-Riley, Peggy O'Neil, and Ann Damiano for their insightful comments on drafts of this research study. Finally, I thank Diep Nguyen for compiling the statistical analysis employed in this study.

[3] Given the pressure our digital footprints are beginning to assert on our lives, privacy advocates may be concerned that the administrator role within My Reviewers may access FYC teachers' comments and grades or those with a mentor role can access their mentees' reviews. In situations like ours in first-year composition programs where courses are being taught by new or inexperienced teachers, this practice seems reasonable and ethical. When we find instructors are providing ineffective commentary, we can provide help and professional training. In contexts where the instructors are tenured professors, this practice is more problematic. In these latter situations the privacy settings can be altered. Given the likelihood of such privacy questions in the future, perhaps the Council of Writing Program Administrators or NCTE/CCCC needs to establish a committee to articulate best practices when it comes to archiving and sharing teacher feedback and scoring. At USF, our initial policy was not to review the instructors' comments, excluding the beginning teachers, unless there were student complaints. A recent analysis of 114,000 comments provided by instructors has challenged us to reconsider whether this is the best policy, however.

Author Note

Joe Moxley, http://joemoxley.org, serves as executive editor and publisher of *Writing Commons*, http://writingcommons.org, the

open-education home for writers. Moxley has published numerous books and articles on assessment, datagogies, qualitative research methods, and commons-based peer production.

References

Anson, C. (2011, December 7). Re: Rubrics and writing assessment. Message posted to WPA-L archives at Writing Program Administration.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.

Astin, A. W. (2011, February 14). In "Academically adrift," data don't back up sweeping claim [Peer commentary by A. W. Astin]. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Academically-Adrift-a/126371/

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science*. Madison, WI: University of Wisconsin Press.

Bazerman, C. (2002). The Case for Writing Studies as a Major Discipline. In G. Olson (Ed.), *The Intellectual Work of Composition*. Southern Illinois University Press.

Beaufort, A. (2007). *College writing and beyond: A new framework for university writing instruction*. Logan, Utah: Utah State University Press.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. Retrieved from http://www.benkler.org/wonchapters.html

Broad, B. (2000). Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English, 35*(2): 213-260.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Logan, Utah: Utah State University.

Carroll, Lee Ann. (2002). *Rehearsing new roles: How college students develop as writers*. Carbondale, IL: Southern Illinois University.

Condon, W. F. (2011, December 7). Re: Rubrics and writing assessment. Message posted to WPA-L archives at Writing Program Administration.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-31). Urbana, 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., & Carlton, S. T. (1961). *Factors in the judgment of writing quality*. Princeton, NJ: Educational Testing Service.

Ebest, S. B. (2005). *Changing the way we teach: Writing and resistance in the training of teaching assistants*. Carbondale, IL: Southern Illinois University Press.

Elbow, P. (1993). Ranking, evaluating, and liking: Sorting out three forms of judgment. *College English, 55*(2), 187-206.

Elbow, P., & Belanoff, P. (2000). *Sharing and responding* (3rd ed.). New York, NY: McGraw-Hill.

Glenn, D. (2011, February 13). Scholars question new book's gloom on education: Doubts are raised about study behind "Academically adrift" [Letter to the editor]. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/scholars-of-education-question/126345

Godshalk, F., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. Princeton, NJ: College Entrance Examination Board.

Graff, G. (2003). *Clueless in academe: How schooling obscures the life of the mind*. New Haven, CT: Yale University.

Graff, G. (2010). Why assessment? *Pedagogy, 10*(1), 153-165.

Haswell, R. H. (1991). *Gaining Ground in College Writing.* Dallas, TX: Southern Methodist University Press.

Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication, 17*(3): 307-352.

Haswell, R. H. (2012, February). Methodologically adrift. *College Composition and Communication, 63*(3), 487-491.

Hosch, B. J. (2010, June). *Time on test, student motivation, and performance on the collegiate learning assessment: Implications for institutional accountability*. Paper presented at the Association for Institutional Research Annual Forum, Chicago, IL. (ERIC Document Reproduction Service No. ED 520481)

Inoue, A. B. (2004). Community-based assessment pedagogy. *Assessing Writing, 9*(3), 208-238.

Lang, S., & Baehr, C. (2012). Data mining: A hybrid methodology for complex and dynamic research. *College Composition and Communication, 64*(1), 172-194.

Langbehn, K., McIntyre, M., & Moxley, J. M. (in press). Using real-time formative assessments to close the assessment loop. In McKee, H., & DeVoss, D. N. (Eds.). *Digital writing assessment*. Edited book under development.

Moxley, J. (2008). Datagogies, writing spaces, and the age of peer production. *Computers and Composition, 25*(2), 182-202.

Moxley, J. (2012). Aggregated assessment and "objectivity 2.0". In M. Piotrowski, C. Mahlow, & R. Dale (Eds.), Linguistic and cognitive aspects of document creation and document engineering: Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2012) (pp. 19-26). Avignon, France.

Nowacek, R. S. (2011). *Agents of integration: Understanding transfer as a rhetorical act*. Carbondale, IL: Southern Illinois University Press.

Pagano, N., Bernhardt, S., Reynolds, D., Williams, M., & McCurrie, M. K. (2008). An inter-institutional model for college writing assessment. *College Composition and Communication, 60*(2), 285-320.

U.S. Department of Education. (2012). *The nation's report card: Writing 2011* (NCES 2012-470). Washington, DC: U.S. Government Printing Office. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf

Vieregge, Q. D., Stedman, K. D., Mitchell, T. J., & Moxley, J. M. (2012). *Agency in the age of peer production*. Urbana, IL: National Council of Teachers of English.

Wardle, E. (2009). Mutt genres and the goal of FYC: Can we help students write the genres of the university? *College Composition and Communication, 60*(4), 765-89.

Wardle, E., & Roozen, K. (2012). Addressing the complexity of writing development: Toward an ecological model of assessment. *Assessing Writing, 17*(2), 106-119.

Wilson, M. (2007). The view from somewhere. *Informative Assessment, 65*(4), 76-80. Retrieved from http://www.ascd.org/publications/educational-leadership/dec07/vol65/num04/The-View-from-Somewhere.aspx

Yancey, Kathleen. (1999) Looking Back as We Look Forward: Historicizing Writing Assessment. *College Composition and Communication, 50*(3), 483-503.