Aggregated Assessment and "Objectivity 2.0"

Joseph M. Moxley
University of South Florida
4202 East Fowler Avenue
Tampa, FL, USA 33620
moxley@usf.edu

Abstract

This essay provides a summary of research related to *My Reviewers*, a web-based application that can be used for teaching and assessment purposes. The essay concludes with speculation about ongoing development efforts, including a social helpfulness algorithm, a badging system, and Natural Language Processing (NLP) features.

1 Introduction

The essay summarizes research that has identified ways *My Reviewers* can be used to:

- integrate formative with summative evaluations, thereby enabling universities and teachers to alter curriculum approaches in real time in response to ongoing assessment information,
- assess students' critical thinking, research, and writing skills—aggregating not a small percentage but all of the marked up documents (in our case about 16,000 evaluations by teachers of students' intermediate and final drafts of essays/semester),
- enable reviewers (teachers and students) to provide more objective feedback, facilitating "Objectivity 2.0," a form of evaluative consensus mediated after extensive crowdsourcing of standards,
- provide conclusive evidence that can be used to compare the efficacy of particular curricular approaches,
- enable students and writing programs to track progress related to specific learning outcomes (from project to project, course to course, year to year),
- inform faculty development and teacher response, and
- create an e-portfolio of students' work that reflects their ongoing progress.

2 What is *My Reviewers*?

My Reviewers is a web-based application that enables students, teachers, and universities to

- aggregate assessment information about students' critical thinking and writing skills,
- mark up PDF documents (with sticky notes, text box notes, drawing tools, etc.),
- grade documents according to a rubric,
- assign and conduct or grade peer reviews. (My Reviewers enables teachers to see at a glance each student's in-text annotations, end-note comments, and rubric scores),
- use a library of comments and resources tailored to address common writing problems, and
- crowdsource comments and resources.

The permissions-based workflow features of *My Reviewers* enable teachers and students to use a rubric and commenting tools to review and grade student writing while protecting student confidentiality behind a Net ID.

My Reviewers is founded on the assumptions that language and learning are social practices, and that students can provide valuable feedback to one another based on their backgrounds as readers and critical thinkers.

By enabling students to track their progress (or lack of progress) according to various evaluative criteria (such as focus, evidence, organization, style, and format), *My Reviewers* clarifies academic expectations and facilitates reflection and awareness of teachers' evaluations and concerns, thereby helping students grow as writers, editors, and collaborators. Furthermore, the pedagogical materials embedded into the tool—videos, explanatory materials, exercises, library of comments with supporting hyperlinks—clarify

Proceedings of the EACL 2012 Workshop on Computational Linguistics and Writing, pages 19–26,
Avignon, France, April 23, 2012.

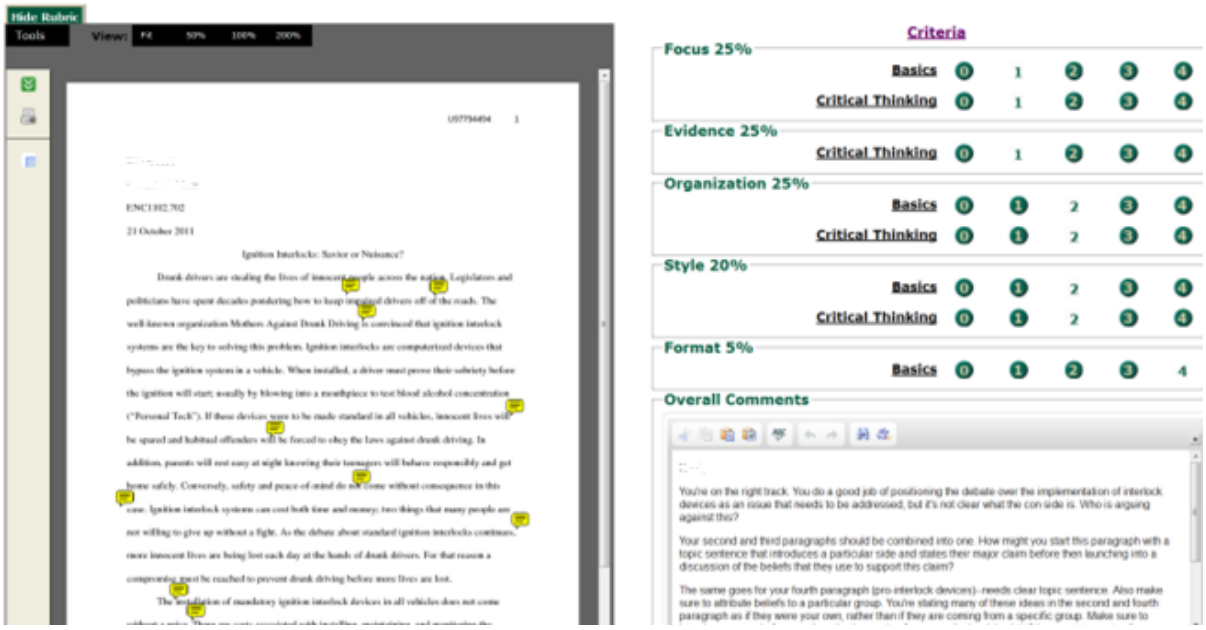

Figure 1: Sample Document Markup and Rubric

grading criteria for both students and teachers. In summary, by aggregating assessment results in innovative new ways, *My Reviewers* reshapes how teachers respond to writing, how students conduct peer reviews, how students track their development as writers and reader feedback, and how universities can conduct assessments of students' development as critical thinkers and writers.

3 Context and Methods

The FYC (First-Year Composition) Program at the University of South Florida is one of the largest writing programs in the U.S, serving approximately 7,500 students in two composition courses each year, ENC 1101 and ENC 1102. Thanks to funding from USF Tech Fee Funds and CTE21, we have piloted use of *My Reviewers* for the past three years, using *My Reviewers* to assess over 30,000 student documents. Last semester (Fall 2011), approximately 70 first-year composition instructors assessed 16,000 essays (including early, intermediate, and final drafts)—not counting student peer reviews. This semester (Spring 2012), we are on course for reviewing another 16,000 essays. The National Council of Teachers of English awarded the FYC Program the *2011-12 CCCC (Conference on College Composition and Communication) Writing Program Certificate of Excellence Award* based in part on its development of *My Reviewers*.

Over the past eight years, our teachers and writing program administrators have crowdsourced a community rubric by employing various peer-production technologies and face-to-face meetings (see Table 1). The early stages of our development process are reported in Vieregge, Stedman, Mitchell, & Moxley's (2012) *Agency in the Age of Peer Production*, an ethnographic monograph published by NCTE's series on Studies in Writing and Rhetoric.

Since moving from a requirement for our instructors to use a printed version of the community rubric to using *My Reviewers,* which enables teachers to view the rubric while grading and associates rubric scores with marked-up texts, we have observed some benefits: While we may have 500 sections of the 1101 and 1102 courses, we want all of these sections to focus on shared outcomes. We have found our use of *My Reviewers* helps ensure students have a more comparable experience than when paper rubrics were used. Back in the days of the printed version of the rubric, at the end of the semester when we surveyed students about usage, about half of our students reported they were unfamiliar with the rubric. One of the advantages of an online tool like *My Reviewers* for universities is that it enables writing program administrators to better ensure instructors and students are keeping up with our shared curriculum. Also, by using a single analytic rubric tool across sections, we can assess progress by student, teacher, section, and rubric criteria.

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 a 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

Table 1: Community Assessment Rubric

As rhetoricians, we understand the value of using rubrics that address the demands of specific rhetorical contexts. When addressing different genres, audiences, disciplines and when using multiple media to remediate texts (Twitter, podcasts, movies, print documents), students clearly benefit from receiving feedback related to conventions in those genres, disciplines, and media. Given this, we clearly understand why Peter Elbow, Chris Anson, William Condon, among other assessment leaders, fault universities for employing a generic rubric like our community rubric to assess texts across projects, genres, courses, media and so on. Like Elbow (2006), Anson (2011), and Condon (2011), we see enormous value in clarifying specific grading

criteria for specific projects, and we understand grading criteria change along with changes in different rhetorical situations. Plus, as compositionists, we understand that writers need different kinds of feedback when they are in different stages of the composing process. Using a rubric like our community rubric early in the writing process can clearly be overkill. There is no point in discussing style, for example, when the writer needs to be told that his or her purpose is unclear or not satisfactory given the assignment specifications. Nonetheless, we have found—as we discuss below—some benefits for using our community rubric to assess multiple projects, even ones that address different audiences, genres, and media.

4 Independent Validation of the Community Rubric by the USF Office of Institutional Effectiveness

While we are currently seeking funding to add administration features that would enable users to write their own rubrics or import rubrics, *My Reviewers* employs a single community rubric (see Table 1) that has been validated by an independent assessment conducted by the Office of Institutional Effectiveness at the University of South Florida in the spring of 2010.

To conduct the assessment, 10 independent scorers reviewed the third/final drafts of 249 students' ENC 1101 Project 2 essays and these same students' ENC 1102 Project 2 essays. The Office of Institutional Effectiveness settled on this odd number—249—because it represented 5% of our total *unique* student head count (4,980 students) for the 2009/2010 academic year. The scorers used the same scoring rubric to evaluate all 498 essays according to eight criteria delineated in our community rubric. Scorers did not provide comments nor did they have access to the markup and grading provided by the students' classroom instructors.

Before the raters scored the randomly chosen student essays, an assessment expert from the Office of Institutional Effectiveness led a brief discussion of the rubric and asked the scorers to read sample essays. He then computed an inter-rater agreement of .93. Confident our scorers understood our rubric and encouraged by our inter-rater reliability, raters subsequently scored the 498 essays over a three-day period.

Naturally, we were pleased to see that our assessment results suggested students were making some progress on all measures of writing and critical thinking, that their 1102 Project 2 scores were higher than their Project 2 scores in 1101, although we were underwhelmed by the degree of improvement. We also were not really surprised that we were able to reach a high level of inter-rater reliability among raters.

However, this study did reveal a counterintuitive and remarkable result: by comparing the rankings of the independent scorers with the rankings of these students' classroom teachers, *we found no statistical difference on seven of the eight rubric criteria*. In other words, when it came to scoring eight criteria, the only difference between the independent scorers and the classroom teachers was "Style (Basics)," a criterion that represents a 5% grade weight when the rubric was used to grade student papers. This discrepancy may suggest that the independent scorers were being more lenient regarding the students' grammatical and stylistic infelicities than the students' classroom teachers.

Overall, the high level of agreement among the classroom teachers and the independent scorers suggests *My Reviewers* (perhaps by clarifying the grading criteria for teachers and students) enables diverse reviewers to mediate a shared evaluation of texts, to reach an unprecedented level of inter-rater reliability among large groups of readers—what we might call "Objectivity 2.0."

In a recent exchange on the Writing Program Administrator Listserv, Chris Anson, this year's Chair of the Conference on College Composition and past president of the Writing Program Administrators writes: "[the] Problem with [generic] rubrics is their usual high level of generalization (which makes them worthless)." In a subsequent co-authored essay, *"Big Rubrics and Weird Genres: The Futility of Using Generic Assessment Tools Across Diverse Instructional Contexts,"* Anson *et. al.* (in press) write: "Put simply, 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."

While we share Anson's preferences for rubrics that are designed to address the particular conventions of specific genres, audiences and media, and while we hope to secure the funding we need to add greater flexibility to *My Reviewers*—so we can better account for different rhetorical situations and media—, our research demonstrates the value and credibility of using a community rubric to assess multiple genres, even

ones that are quite distinct, such as the personal narrative essays versus third-person based research reports. Perhaps our results suggest that the eight criteria defined by our rubric are generalizable enough across disciplines, genres, and media that university faculty can recognize them and employ them in meaningful ways to reach Objectivity 2.0.

To be completely frank, we are somewhat astounded by the inter-rater reliability we have been able to achieve among such diverse readers, and we wonder whether a rubric such as our community rubric can be used meaningfully to overcome the "courseocentrism" that Gerald Graff (2010) has described as undermining education in the U.S. Perhaps a tool such as *My Reviewers* can be used to leverage communication across departments, perhaps general-education wide, to address the common characteristics of academic prose that faculty across disciplines value.

5 Assess Undergraduate Learning

Richard Arum and Josipa Roksa have received worldwide attention for their evidence and argument in *Academically Adrift* (2011) that undergraduates fail to learn much despite their coursework. In contrast, by comparing students' scores from project to project, we have been able to demonstrate students' development as writers, researchers, and critical thinkers. Note, for example, our evidence, shown in Figure 2, of student development over one academic semester—based not on a small sample size but on *all* students in ENC 1102 that semester.

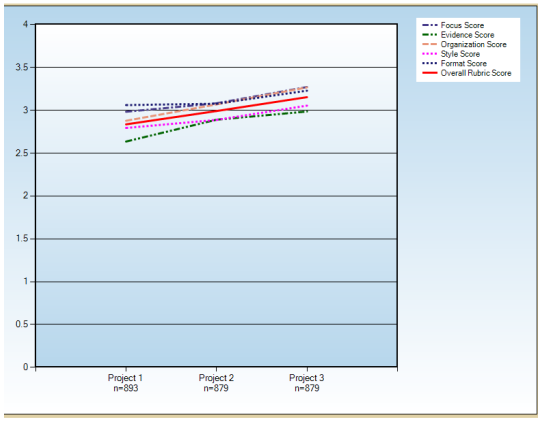

Figure 2: 1102 Final Project Scores

6 Make Evidence-Based Curriculum Changes

As any seasoned teacher or administrator knows, not all curricular materials are equivalent. On occasion, students perform poorly not because of a lack of innate inability but because of poor curricular planning on the part of the teachers (e.g., inadequate scaffolding of projects). Figure 3 illustrates ways *My Reviewers* can be used to improve the curriculum in light of evidence—illustrating ways assessment results can be used to inform curriculum changes. In this example, program administrators made changes to the historiography project (Project 2) from the Spring 2010 semester, and, subsequently, in the Fall 2011 semester students scored significantly better on most measures (Langbehn, McIntyre, Moxley, 2012).

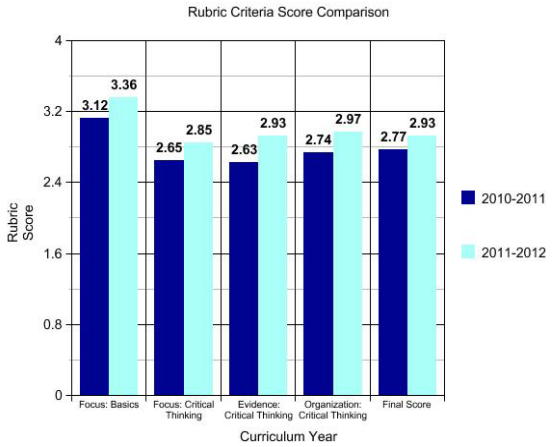

Figure 3: Comparison of Project 2 for the Spring 2010 vs. Fall 2011 Semesters

7 Compare Alternative Curricular Approaches

Use of a community rubric across genres, courses and disciplines can also be used to chart student progress, or lack of progress, or to indicate distinctions between the levels of difficulty imposed by unique projects/genres. On occasion, the lack of student success can be linked to issues pertaining to curriculum design as opposed to a particular student deficit. Figure 4 shows the comparison of student scores in two alternative courses, taken in succession by students at our university—results that suggest we need to once again rethink our curriculum for 1101 despite our intuition that the course was well designed and well received:

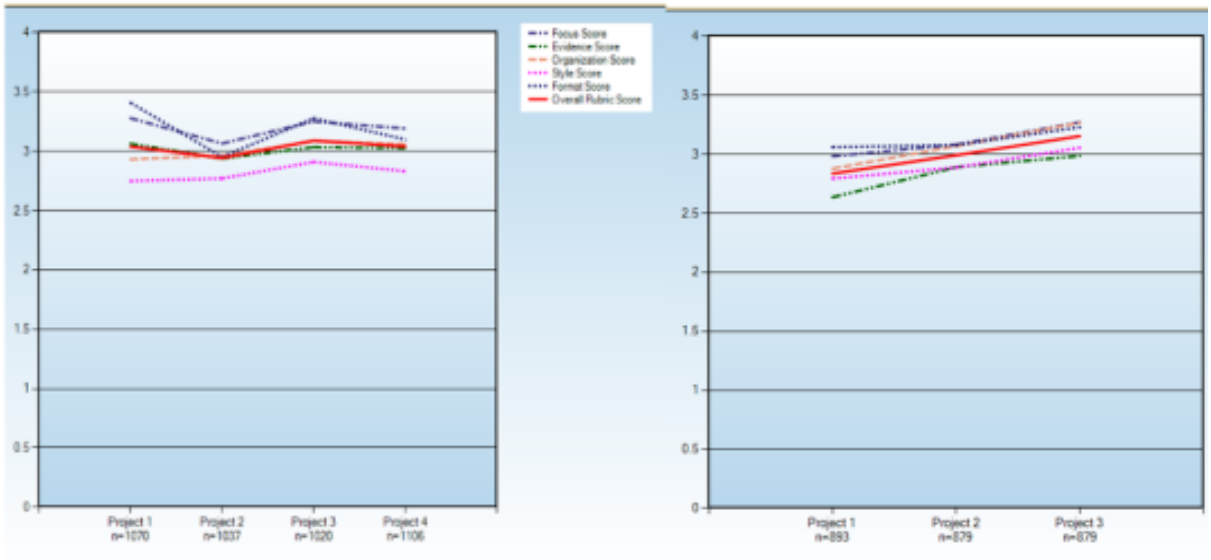

Figure 4: 1101 (left) vs. 1102 Final Project Results

8 Develop and Compare New Models for Teaching and Learning

Writing programs can use tools such as *My Reviewers* to compare alternative curriculums. We are currently providing three alternative approaches to teaching writing in university settings—the traditional approach, where students meet three hours each week in class; an online model; and a collaborative model, which requires students to use *My Reviewers* to conduct two cycles of peer review and two cycles of teacher feedback—as illustrated partially in Figure 5.

9 NLP Features Under Development

We are currently implementing a library of comments, which we developed by analyzing approximately 30,000 annotations and 20,000 endnotes; we are in the process of developing resources to help students better understand teacher and peer comments.

We are seeking additional funding to develop an algorithm and badging system to inspire more effective peer-review. By enabling students to earn badges according to the quality of their feedback, as measured by their peers and students, we are hoping to provide a further incentive for quality feedback. We would like to tie the badges to the number of substantive and editorial critiques that the document authors account for when revising, by endorsements by teachers for peer feedback, and by overall rankings of peer reviews.

Eventually we hope to add NLP (Natural Language Processing) tools that identify repeated patterns of error—as identified by past and present teachers who have used the tool. For example, students could be informed when they have received similar feedback in the past, and they could be offered hyperlinks back to past, similar comments. We can imagine features that highlight for teachers common comments on specific sets of papers or projects. Perhaps OER (Open Education Resources) such as Writing Commons, http://writingcommons.org, could be suggested as teachers and peers make comments.

10 Conclusions

In his seminal work, *The Wealth of Networks,* Yochai Benkler wisely remarks,

> Different technologies make different kinds of human action and interaction easier or harder to perform. All other things being equal, things that are easier to do are more likely to be done, and things that are harder to do are less likely to be done. (17)

My Reviewers, and other tools like it that are in development, shatter pedagogical practices by making it easier to provide comments, easier to organize and grade peer reviews, and easier to conduct assessments based on whole populations rather than randomly selected groups. The Learning Analytics embedded in tools like *My Reviewers* can empower students, teachers, and administrators in meaningful ways.

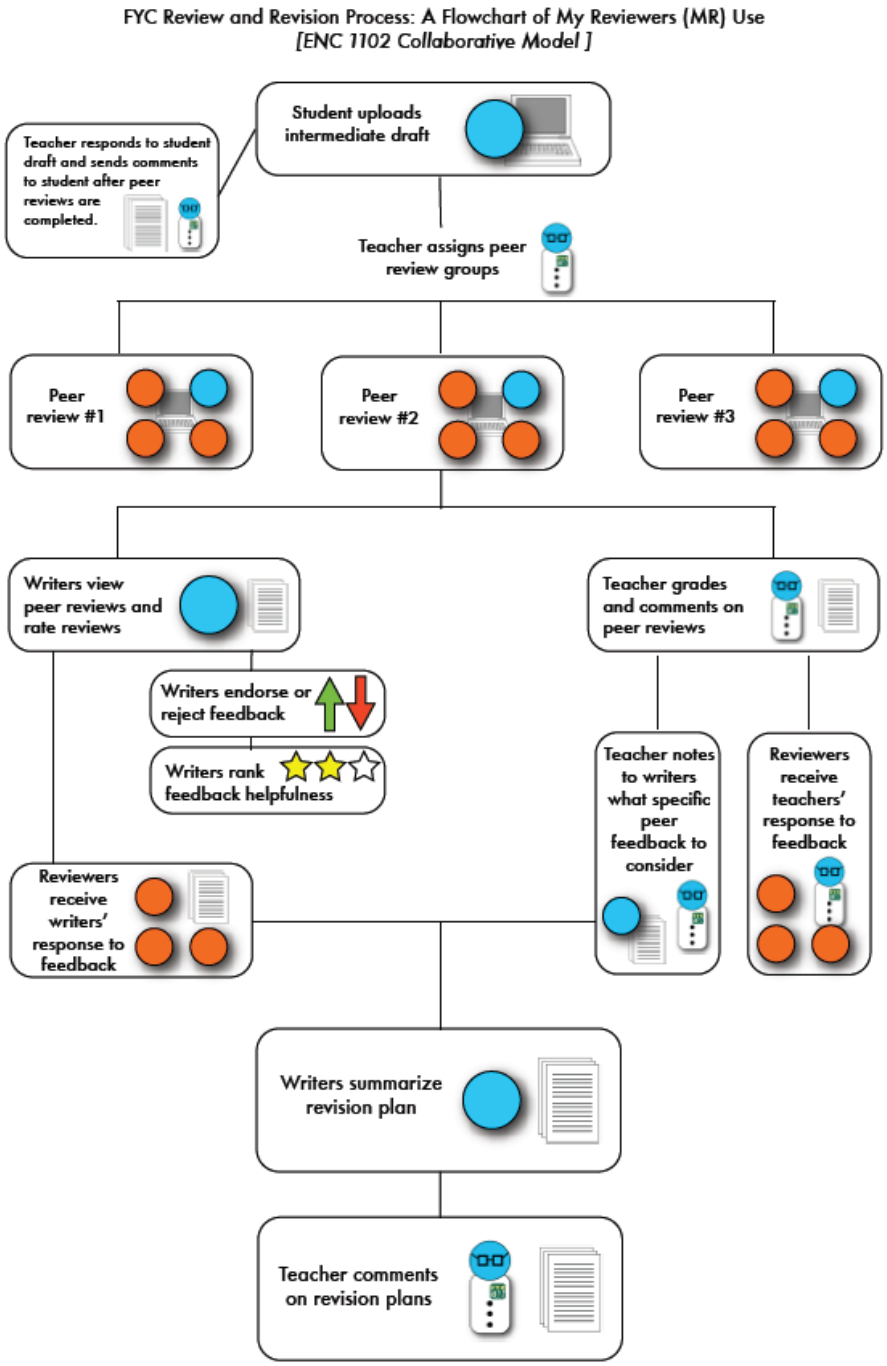

Figure 5: Cycle 1 for Peer Review Process

Acknowledgments

Project Development has been a deeply collaborative effort. Terry Beavers, Mike Shuman, and I—the chief architects of *My Reviewers*—have benefitted from the contributions of many colleagues. We thank Michelle Flanagan, for her ongoing development work; Dianne Donnelly; Hunt Hawkins; Janet Moore; Steve RiCharde; Dianne Williams; Nancy Serrano, Megan McIntyre; Nancy Lewis; Brianna Jerman; Erin Trauth. Finally, we thank the University of South Florida Technology Fee Grant Program and the Center for 21st Century Teaching Excellence for funding our project.

References

Chris M. Anson. 2011. Re: Rubrics and writing assessment. In WPA-L Archives. Council of Writing Program Administrators. Message posted to http://wpacouncil.org/wpa-l

Chris M. Anson, Deanna P. Dannels, Pamela Flash, & A.L.H. Gaffney. In press. Big Rubrics and Weird Genres: The Futility of Using Generic Assessment Tools Across Diverse Instructional Contexts. Journal of Writing Assessment.

Richard Arum & Josipa Roksa. 2011. Academically Adrift. University of Chicago Press, Chicago.

Yochai Benkler. 2006. The Wealth of Networks. Yale University Press, New Haven and London.

William F. Condon. 2011. Re: Rubrics and writing assessment. In WPA-L Archives. Council of Writing Program Administrators. Message posted to http://wpacouncil.org/wpa-l

Peter Elbow. 2006. Do We Need a Single Standard of Value for Institutional Assessment? An Essay Response to Asao Inoue's 'Community-Based Assessment Pedagogy'. Assessing Writing, 11:81–99.

Gerald Graff. 2010. Why Assessment? Pedagogy, 12(1):153-165.

Karen Langbehn, Megan McIntyre & Joseph Moxley. Under review. Using Real-Time Formative Assessments to Close the Assessment Loop. In Heidi McKee & Danielle Nicole DeVoss (Eds.), Digital Writing Assessment.

Quentin Vieregge, Kyle Stedman, Taylor Mitchell, and Joseph Moxley.. In press. Agency in the Age of Peer Production. Studies in Writing and Rhetoric Series. National Council of Teachers of English, Urbana, IL.