On Keeping Score: Instructors' vs. Students' Rubric Ratings of 46,689 Essays

Joseph M. Moxley and David Eubanks

Abstract

This study investigates the efficacy of having first-year composition students score one another's intermediate drafts of essays using a five-trait rubric across 482 sections of two introductory composition courses (ENC1101 and ENC1102).[1] *This study analyzes 46,689 reviews, which consisted of 16,312 reviews conducted by instructors and 30,377 reviews conducted by students. The papers, typically between 1,000 and 1,500 words each, were written by students over the course of seven semesters at University of South Florida, a state university in the United States. We found low to modest correlations between peer ratings and instructor ratings on individual assignments. On average, peers assigned higher ratings than instructors, yet, over time, students' scores were more highly correlated with instructors' scores. The average differences in ratings between the students and instructors were smallest for* Focus *and* Format *and greatest for* Evidence. *Students who received higher ratings on their own writing from instructors provided scores that had a broader range of scores and were more highly correlated with instructors' scores than students who received lower scores from instructors. Generally, peers had a smaller rating variance of scores than instructors.*

While a good many pedagogical essays have been published regarding best classroom practices for conducting peer review, surprisingly little quantitative, replicable, aggregated, data-driven (RAD) research has been conducted on peer review in the discipline of Writing Studies. Regarding the paucity of empirical research in NCTE journals, Richard Haswell concluded that "peer critique seems to be one of the least studied of practices now very common in college writing classrooms" (211). One overlooked question is whether or not it is worthwhile to have students assess other stu-

dents using a rubric. While a few studies have compared instructor scores to student scores to interrogate the validity of peer review, to reach the effects of training on peer reviews, or to investigate gender bias in peer review, these studies have been limited by small sample sizes. To our knowledge, only one of these studies questioned if the correlation between instructor and student scores changes over time or whether reviewers are more likely to reach higher levels of correlation for particular rubric criteria.

At the first-year composition program (FYC) at the University of South Florida, we are especially interested in this question because we would like to know whether it is a worthwhile practice to require students to score one another's intermediate drafts. We also wanted to evaluate whether our practice of using one generic rubric across sections of two composition courses was working well. The FYC rubric had been developed via a crowdsourcing process conducted by writing instructors, writing program administrators, and USF's Office of Institutional Effectiveness (Vierrege et al.). Past research addressed the internal reliability of the numeric rubric (Moxley, "Aggregated Assessment"); illustrated ways writing program administrators have deployed the numeric rubric to make real-time, evidence-based curriculum enhancements (Langbehn et al.; Moxley, "Big Data"); and analyzed instructors' and students' use of rubric criteria in 118,611 comments made on 17,433 student essays (Dixon and Moxley).[2] While we identified benefits to having instructors and students use the same rubric across genres and courses (Moxley, "Big Data"; Anson et al., "Theorizing Community"), we were troubled by how our policies contradicted modern assessment theory—that is, the assumption that every rhetorical situation warrants a unique rubric (Anson et al., "Big Rubrics"). We wondered whether we could better serve our students by using different rubrics for different genres and purposes. Initially, My Reviewers, the tool we use to markup student documents and to conduct peer reviews and team projects, didn't permit customizable rubrics. But after that particular technical obstacle was overcome, we wondered if we should revisit the possibility of diverse, multiple, project-specific rubrics for peer reviews (see http://myreviewers.com for software details). When we broached this possibility with our instructors and colleagues, we faced some resistance, so we wanted to research the efficacy of our existing measures before considering the move to multiple rubrics or discontinuing the practice of asking students to score as well as comment on peers' works.[3]

In summary, this study uses corpus methods to investigate the efficacy of having students score one another's essays as opposed to solely providing textual comments. In particular, we analyze the rubric scores provided by instructors and students who used the numeric version of the FYC rubric to

assess intermediate drafts—that is, 16,312 reviews provided by 107 instructors and 30,377 peer reviews provided by 5,857 students—for the three major projects in two introductory composition courses between the spring 2012 and spring 2014 semesters.[4] Through our exploration, we sought to mainly explore these questions:

1. Would students score similarly to their instructors?
2. Would the correlation between students' scores with instructors' scores improve over time (within a class and over the year in both composition courses) and why might that be the case?
3. Is peer scoring an effective pedagogical practice?
4. What changes to our curriculum—including our implementation of the rubric, the design of the rubric, the training of students to prepare them for peer review—are suggested by study results?

Literature Review

In the discipline of Writing Studies, empirical work on peer review has been sparse, especially given the popularity of this pedagogical practice. So little empirical work has been published in the flagstaff publications of the Conference on College Composition and Communication and the National Council of Teachers of English, in fact, that Richard Haswell contended in 2005 that Writing Studies was at war with "empirical inquiry, laboratory studies, data gathering, experimental investigation, formal research, hard research, and sometimes just research" (200). Outside of Writing Studies, however, in the broader assessment and education literature, research on peer review has been fairly robust. In his literature review spanning 1980 to 1996, Keith Topping used the search terms "*peer assessment*, *peer marking*, *peer correction*, *peer rating*, *peer feedback*, *peer review*, and *peer appraisal* (together with university, college, and higher education") ("Peer Assessment" 250) to find 109 publications that focused on peer review in the Social Science Citation Index, Dissertation Abstracts International, and ERIC databases. Of those 109 studies, 67 articles "included outcome data gathered in an orderly research process" (250). Regarding the validity and reliability of peers' assessments, Topping concluded, "Peer assessment of writing and peer assessment using marks, grades, and tests have shown positive formative effects on student achievement and attitudes; these effects are as good as or better than the effects of teacher assessment" (249). Interestingly, in the 25 studies that compared teachers' marks or grades with stu-

dents', researchers reported high reliability between teachers and students in 18 (72%) of the studies (257).

Using Topping's same search terms in ISI Web of Science (formerly the Social Science Citation Index), we found 23 empirical studies had been published on peer review between 1997 and July of 2014. Continuing the trend identified by Haswell, all of these RAD studies were published in non-NCTE journals.[5] These research studies address a range of topics (e.g., peer review and gender, experiences of L2 students, attitudes toward peer review, effects of training, and validity of peer reviews versus self-assessment or teacher assessment) and methods (survey, observational, quasi-experimental, and meta-analysis). One fairly robust theme in the literature is the question of gender bias in peer reviews, and these studies have occasionally compared peers' and instructors scores by gender (Tucker; Falchikov and Magin). Several studies have compared students' and instructors' scores on papers to interrogate the validity of peer reviews (Falchikov and Boud; Falchikov and Goldfinch). Liu and Lu found that after receiving training in peer review strategies, the correlation between students' scores and instructors' scores significantly increased. Esfandiari and Myford compared the ratings on an eight-point analytical scale provided by 194 assessors on 188 essays and found that teachers were most critical, then peer assessors, and then self assessors. When Liang and Tsai compared self, peer, and expert assessments on a four-point analytic scale (Knowledge, Suitability, Correctness, and Creativity) to assess biology reports written by 47 students, they found good consistency between students and experts and found that inter-rater agreement improved over time.

Perhaps due to the limitations of traditional data collection techniques, one important limitation of past research on peer review in general has been that it has been primarily constrained by small sample sizes.[6] Excluding Takeda and Homberg's study, which analyzed the peer reviews of 1,001 British students, and Tucker's 2014 study of gender in peer review with a sample of 1,523 students, no large-scale empirical work has been conducted on peer review. Instead, past empirical investigations have been limited—sample sizes are typically fewer than 50 (Khonbi and Sadeghi; Liu and Lee; Liu and Li; Lundstrom and Baker). Only five studies have worked with samples of more than 200 students (see table 1).

Table 1
Largest quantitative sampling plan for studies on peer review, 1997-2014

N (students)	Study	Method	Finding(s)
1,523	Tucker (2014)	Group mean comparison	Absence of gender bias in peer assessments Women received significantly higher peer ratings than men ($p < .05$)
1,001	Takeda & Homberg (2014)	Individual level comparison Group mean comparison	Gender balanced groups show notably lower variation in self- and peer-assessment scores Enhanced collaboration between students in gender balanced groups
300	Boase-Jelinek, Parker, & Herrington (2013)	Individual level comparison	Students did not interpret the rubric in a similar manner as their tutor.
211	Patchan, Schunn, & Clark (2011)	Group mean comparison Correlation	Students' drafts were of higher quality when written for peers than when written for their teacher's assistant Students provided more detailed reviewer comments than teaching assistants However, between student and TA reviewers, only moderate differences were found in final draft scores
208	Crossman & Kite (2012)	Individual level comparison	Use of rubrics with peer reviews resulted in improved quality of students' papers between the initial and final drafts.

WPA: Writing Program Administration, Volume 39, Number 2, Spring 2016

Research Design

In this study, we use statistical measures to compare instructors' scores on intermediate drafts with students' scores on these same drafts. Instructors and students used the numeric version of the FYC rubric to score intermediate and final drafts.

Setting

Over a three-year period, 128 instructors used My Reviewers to assign students to peer review groups, typically two or three students per group. The papers students were asked to review constituted the primary coursework/grades for two composition courses (ENC1101 and ENC1102): annotated bibliographies, literature reviews, analytic essays, historiographies, Rogerian arguments, remediations, and arguments for social justice. We offer three versions of ENC1101: a completely online model, a traditional model wherein students meet in classes (with enrollments capped at 25/class during the fall and 22/class during the spring), and a flipped model, wherein students meet for about an hour in large groups each week and then meet two hours either one-on-one with instructors or in small groups. We offer two versions of ENC1102: traditional and online. While the nature of these courses changed somewhat from year to year, the three projects in each course were designed to be increasingly more difficult, moving from summary to synthesis to argument (see http://fyc.usf.edu/~hosted for project details). All students were required to write three drafts of each project: 1) a preliminary draft that the instructor and student discussed, typically in a one-on-one conference; 2) an intermediate draft that the instructor and students reviewed independently; and 3) a final draft that only the instructor reviewed. Regardless of which type of class they were enrolled in, all students conducted their peer reviews anonymously online, using My Reviewers. In total, during this time period, first-year composition instructors reviewed 16,312 intermediate drafts and students reviewed 30,377 intermediate drafts using the numeric version of the FYC rubric.

Given that we are a fairly large community comprised primarily of graduate students who have disparate teaching schedules and that our curriculum already takes 10 hours to teach each week—the number of hours for which we pay our graduate students each week—it would be unreasonable to ask for more time from our instructors. We have found it nearly impossible to host grade-and-comment norming sessions although we recognize the value of such sessions in terms of facilitating stronger inter-rater reliability among instructors in our program. Beyond approximately 24 hours of training in our yearly fall orientation, grade norming is limited to the new

graduate students who enter our program each year and who are required to take a semester-long practicum that meets weekly. To help facilitate a shared language for assessment and response, therefore, we have provided a variety of peer-review videos and sample document markups in the three ebooks we have developed for our students.

All students used My Reviewers to conduct their peer reviews. Generally, our instructors graded and commented on intermediate drafts at the same time students conducted peer reviews on the same intermediate drafts. However, a few instructors required students to rewrite their projects after the peers' reviews and before they graded or responded to them. In addition, there were some additional variances regarding ways instructors used My Reviewers. For example, some instructors simply assigned the peer reviews and neglected to grade them; others graded the peer reviews but did not discuss them in class; some broke the anonymity of the peer reviews as experienced at My Reviewers and asked students to conduct follow up face-to-face sessions in class after their online peer reviews; some discussed peer review in class meetings, as recommended by our sample detailed schedule, but others did not.

When using My Reviewers, instructors and students have a range of features that they may or may not utilize. For example, students and instructors may use the .pdf-markup tools to write comments and draw on the papers; write endnotes that explained their in-text comments; and place Community Comments on one another's documents, which are hyperlinks out to an article, video, and *Try It! Exercises* about the comments.[7] Within the My Reviewers document-workflow system, instructors may view from a single page all aggregated sticky notes, endnotes, Community Comments, and rubric scores each student provided on assigned peer reviews (see figure 1). Below that information, instructors may grade peer reviews and write a note to the student regarding his or her review. Instructors may also double click to see each peer review from the student view.[8] Adoption of these features has been varied across instructors.

Instrumentation

The community rubric all instructors and students used during the conduct of this investigation contains five broad categories—*Focus*, *Evidence*, *Organization, Style,* and *Format.* Three of the rubric criteria—*Focus*, *Organization*, and *Style*—contain two subcategories: *Basics* and *Critical Thinking*. *Evidence* includes the *Critical Thinking* subcategory whereas *Format* includes *Basics*. The *Basics* subcategory focuses on language conventions such as grammar, mechanics, and punctuation, while the *Critical Thinking* subcategory identifies global rhetorical concerns.

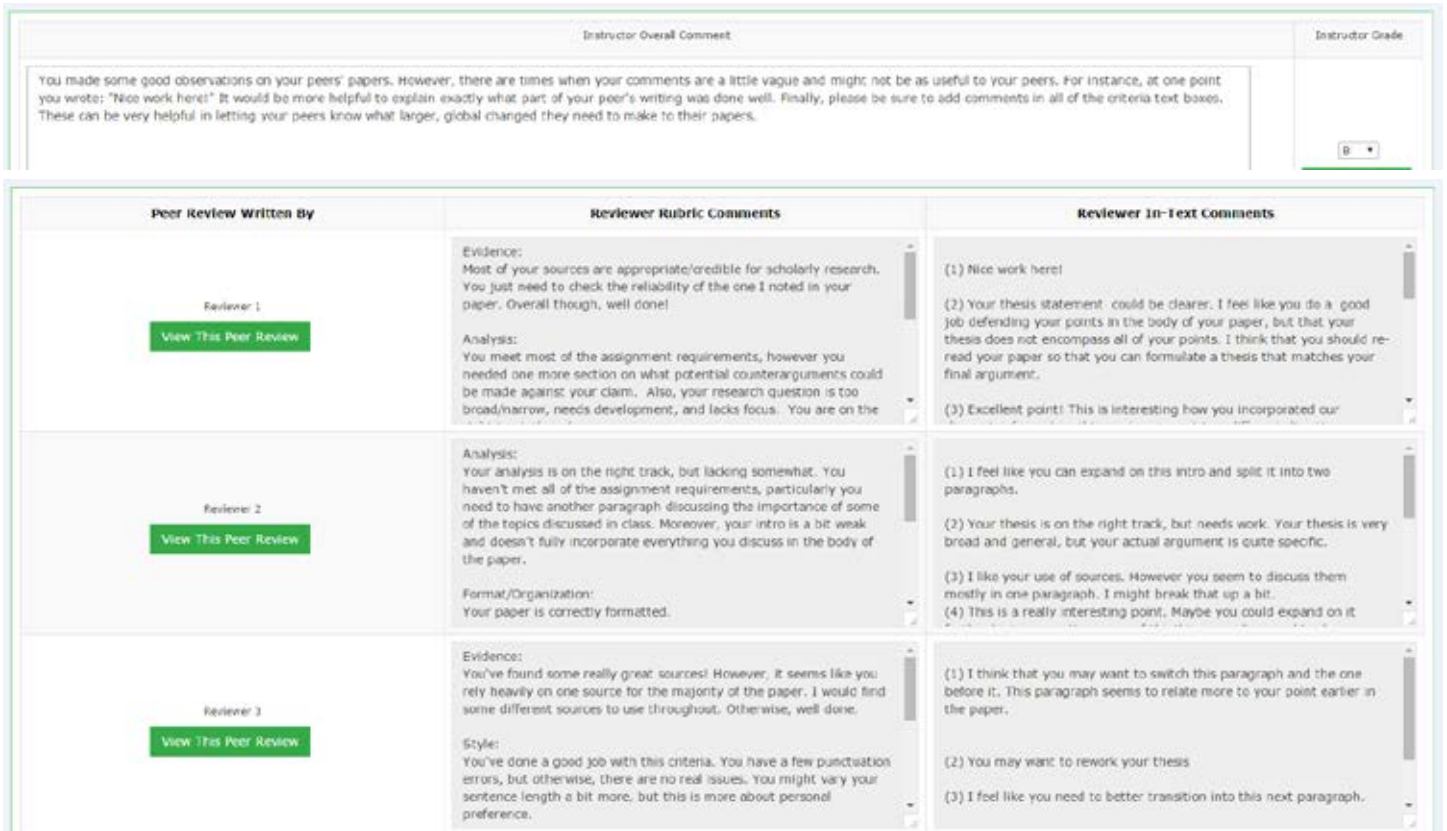

Fig. 1. Sample peer review and instructor grade of a review.

During the early part of the study—spring 2012 through fall 2013—instructors could choose between two versions of the community rubric when assigning peer review: the numeric rubric, which requires students to score rubric criteria on a five-point scale, and the discuss rubric, which requires students to write textual comments regarding these criteria rather than scores. When presented a choice, instructors have favored the numeric version of the FYC rubric over the discuss version, perhaps because the numeric version is the default rubric defined by the My Reviewers system or perhaps because students who are required to take the composition classes are often grade focused and prefer to know where they stand grade-wise at the intermediate draft stage. During the spring 2014, instructors were only presented with the numeric rubric option (see table 2).

Research Limitations

Beyond the rubric scores, students and instructors provided on intermediate drafts reported on in this study, reviewers provided written comments, including .pdf sticky notes and drawings and text notes on top of students' papers, rubric-based comments, Community Comments, summary notes, and revision plans. However, given scope limitations, this study does not provide an analysis of these lexical comments nor does it compare improvement from intermediate to final projects. In addition, given the limitation of the IRB protocol that we followed for this study, we do not analyze students' scoring by students' gender, SAT scores, college or high school grades, ethnicity, or First Time in College status.[9] Finally, it is important to note that this is a purely observational study, and, following Schneider et al., we make no claims about causality.

Results

This study reveals large differences between instructors' and students' scores on intermediate drafts written as the primary coursework for two introductory composition courses. Generally speaking, students score higher than instructors, particularly in the first project in ENC1101, although over time the correlation between students' and instructors' scores improves.

What Is the Agreement between Instructor- and Peer-Assigned Ratings?

Assuming instructors' reviews represent the gold standard, it follows that close association between peer and instructor scores indicates desirable metacognitive skills on the part of students with regard to assessing writing and presumably leads to better writers.

Table 2
The Common Rubric for First-Year Composition

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

Table 2 cont.
The Common Rubric for First-Year Composition

Criteria	Level	Emerging 0	1	Developing 2	3	Mastering 4
	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

WPA: Writing Program Administration, Volume 39, Number 2, Spring 2016

We began our analysis by examining the inter-rater properties between peers and instructors. Inter-rater reliability is an assessment of how well two or more raters implicitly rank each observation relative to the others (do they rise and fall together?). On average, individual raters may tend to assign higher or lower scores than others and still agree on ranking, which we measure with correlation of average scores across each peer-reviewed paper (see fig. 2).

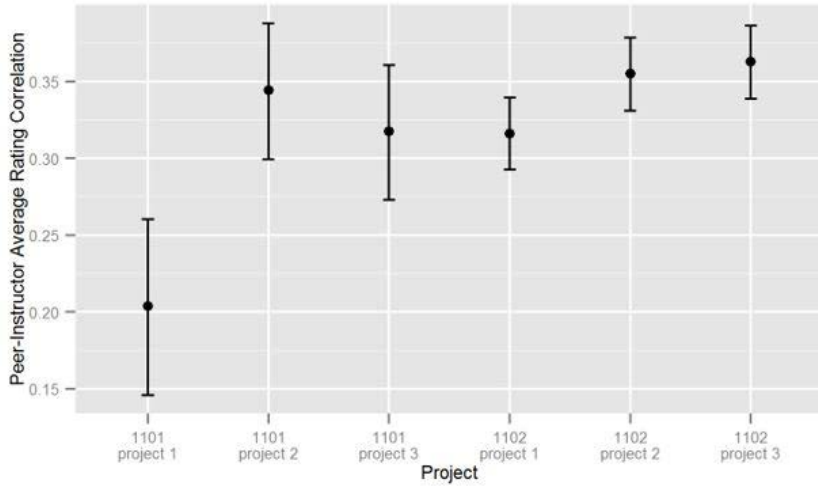

Fig. 2. Peer review instructor average rating by project.

The correlations in figure 2 are between the average score over the eight sub-scores for peer ratings and instructor ratings on samples sizes of greater than 1000 for ENC1101 and greater than 5000 for ENC1102. The dot in the middle is the estimated correlation, and the upper and lower range shown with the bars are 95% confidence intervals. The correlations are low to modest, but in addition to showing that there is some relationship between student rater and instructor on individual papers, the sequence shows that raters improve this correspondence after the first project in ENC1101. When we compare the intervals graphed in figure 2, we can see that the differences between correlations show a large gap between the first project of ENC1101 and the subsequent projects ($p < .002$, $z > 3.1$ for each of these comparisons). The numerical differences between the subsequent projects are not statistically significant. This suggests that there probably is some unique aspect of the first assignment in ENC1101 that produces lower student-instructor rating agreement. For example, perhaps the first assignment in ENC1101 teaches students how to better conform to rating norms.

We next visually compared the average ratings of peers and instructors by graphing the two together. The average score can range from zero to four, and if the scores matched perfectly, the graph would just be points marked by circles (one per student) that fall on the diagonal line from zero to four. We already know from the correlation analysis that the agreement isn't perfect, so realistically we expect a cloud of points, and the shape of the cloud may have some interest to us (see fig. 3).

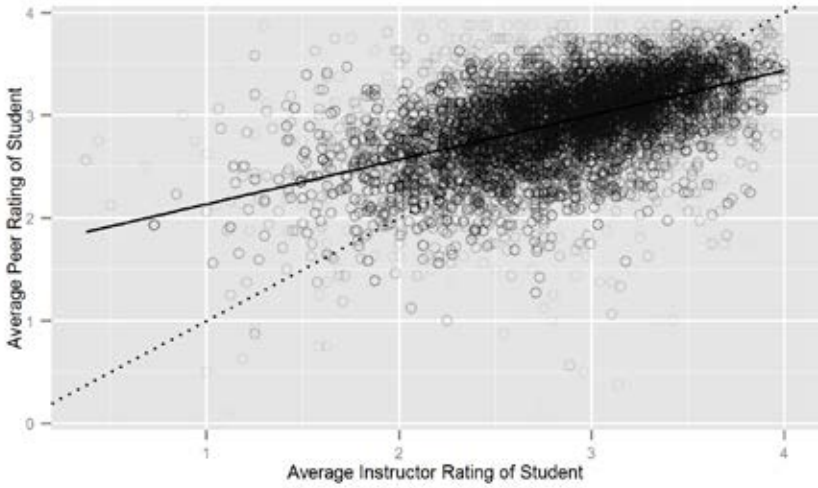

Fig. 3. Scatterplot comparing instructor ratings to peer ratings.

The scatterplot in figure 3 has a circular mark to denote each student's average scores assigned by the instructor (horizontal scale) and peers (vertical scale). We can see that most of these points fall between an average of 2 and about 3.8 in both cases, which means that zeros and ones are being reserved for exceptionally poor (or incomplete) papers. This is good news for the students, whose grades depend on these scores, but it is less optimal from a measurement point of view because it squashes the measurements together and makes it harder to distinguish cases. This undoubtedly contributes to the low correlation between peer and instructor scores.

The dotted diagonal line shows where peers and instructors agree on the score for a given student. There is substantial disagreement, but a linear regression model can find the straight line through the cloud of points that fits the data as well as any straight line can do. The result is the solid line on the graph, and the technical details are as follows: R^2 =.21, $F(1, 5776)$ =1580 (p<.001). Comparing that to the dotted perfect match reference line, we can see that peers have a tendency to rate higher than instructors for

average ratings less than 3 and rate somewhat lower than instructors for ratings greater than 3.

An analysis of inter-rater agreement shows that peer raters have a hard time distinguishing between the 3 versus 4 rating on the scale. That is, the distinction appears to be statistically random. This is less true for the 1 versus 2 ratings, implying that students do have some skill in making quality distinctions, but that this ability is commensurate with their own development as a writer. The good news is that peers had a relatively easy time distinguishing the 1s from 4s, so their powers of discrimination do exist, but perhaps not at the finest level of the rubric.

Taken together, these findings suggest several ways to generate higher quality peer ratings. One is to combine training with pedagogy (e.g. using a rubric as a teaching tool) and more narrowly focus on specific aspects of writing within a given assignment. For example, instead of using the generic rubric, a simpler one with narrower characteristics may allow inexpert raters to more easily identify the traits on the rubric. Pedagogical use could also extend to training assignments where students rate pre-scored samples to see if they get the correct score. Finally, the issue of which end of the rubric to concentrate on seems to emerge from these findings. That is, do we build on the seeming natural ability of students to distinguish poor from average work (or whatever we call the low end of the scale), or do we work harder to simultaneously develop their writing with the metacognitive skills required to distinguish good from excellent work? This is, in itself, an interesting research question. We believe it is reasonable to assume that being a good critic is actually harder than being a good writer, and we might want to plan the curricular path of peer review to incorporate this lag. It also raises the question of whether peer reviews could do more harm than good in responding to good or excellent work. Perhaps, after all, students should not be asked to score papers but instead provide lexical comments, although without more research we cannot be sure about the veracity of their lexical comments.

Finally, regarding classroom implications, it is important to note that these results suggest the quality of reviews is more likely to be linked to the strength of the writers in each group rather than the total number of students in a group, given stronger writers had a broader spread of scores and a higher correlation of scores in relation to their instructors' scores on their papers. Hence, writing instructors should give some thought to the quality of writers in a group and not just the size of the group.

How Do Instructors' Scores by Rubric Criteria Compare with Students' Scores by Rubric Criteria on Intermediate Drafts?

In figure 3, we saw that peers tend to rate higher than instructors when the average is less than 3 and lower than instructors otherwise. We now investigate the differences for each of the eight rubric traits. Close association between the peer and instructor ratings indicates desirable metacognitive skills on the part of students with regard to assessing writing and that presumably leads to better writers. Therefore, the differences between peer and instructor ratings may tell us relative strengths and weaknesses of particular traits within the rubric (see fig. 4).

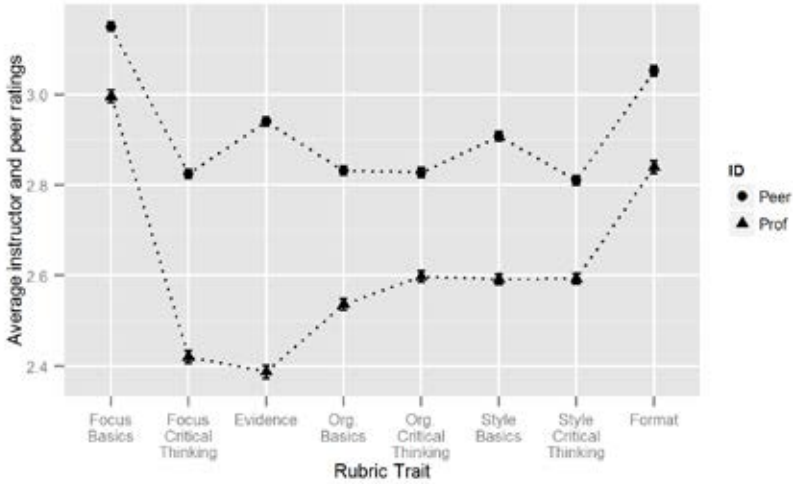

Fig. 4. Average instructor and peer ratings by rubric trait.

Figure 4 shows that, on average, peer reviewers rate higher than instructors on every trait. We can see from the confidence intervals that the differences are all statistically significant (because the error bars are not overlapping), and using Welch's t-test of difference of means, each difference has $p<$ .001. The difference between peer and instructor is least for *Focus 1* (which assesses the degree to which the text addresses the assignment prompt) and *Format*, which are arguably the easiest traits to assess. In contrast, *Focus: Critical Thinking* and *Evidence* have the largest gaps. The former is not surprising since it is probably the most complex of all the traits that students are asked to assess. *Evidence* seems much easier to identify. Here again, a developmental model of writing and metacognition would be useful. We make the tentative, but we think reasonable, conclusion that student and instructor ratings are best aligned when the trait being assessed is simple and therefore easier for a novice to identify.

Do Better Writers Score More Similarly to Instructors than Weaker Writers?

We calculated the overall score received by each student from his or her instructor as a proxy for expressed writing ability and divided the students into quartiles from lowest to highest. Within each of these ranks, we computed the correlation of scores those students gave as peers to the scores assigned by instructors on the same papers, using the average of the eight rubric subscores for comparison. The lowest ability peer raters (as measured by the scores on their own papers) have a .25 to .31 correlation with their instructors when rating the same papers. In contrast, the highest ability peer raters have a confidence interval of .34 to .40 (all correlations are different from zero with $p<.001$). This difference between the lowest and highest quartile is significant ($p < .001$). This trend suggests that—as one would expect—students who receive better ratings on their own writing from their teachers make better peer reviewers, assuming, again, that the instructors' score represents the gold standard. The correlations are low in absolute terms, and we have already seen that the peer ratings have several validity issues, but the large number of samples allows us to detect this slight tendency of increased alignment between instructors and the best writers. There is one other element of this agreement that is worth mentioning: Statistically, correlations tend to increase as the variance, or *spread*, of ratings increases. We already noticed that the limited use of the full range of the 0-4 scale presents research problems, and the same is true of individual raters. To illustrate, imagine a rater who only ever assigns a 3. In effect, no information is transmitted with these ratings because they cannot distinguish between levels of quality. Overall, peers had a smaller rating variance (median = .24) than instructors (median = .42), meaning that peer ratings provide less information for either feedback or assessment purposes than do instructors, and the disparity serves to lower the correlations between peers and instructors. Assigning a wider spread of ratings is presumably a trainable skill and could even be enforced in an online system (imagine having only a limited number of 4s to hand out, for example). At the least, the awareness of score spread can be reinforced by reporting it to raters as they rate. We envision this conscious discrimination and feedback as a general teamwork skill with applications in many types of collaboration.

Do We Really Need Eight Rubric Traits?

Peer reviewers could understandably have trouble distinguishing between the eight distinct ratings they assign, each of which is supposed to assess a different aspect of a paper. In fact, we cannot expect even the best raters to cleanly distinguish between them. We can imagine a paper that has perfect

organization but is utterly lacking in style or a paper that is formatted to perfection but lacks an evidentiary basis. However, when we mix together eight of these traits (or dimensions), it is to be expected that relationships between them become evident. One way to detect that is to simply look at the correlations between scores assigned to the eight individual traits (see table 3).

Table 3
Correlation Matrix of Rubric Traits for Peer Ratings (n = 30,377)

	1	2	3	4	5	6	7	8
1. Focus (Basics)	-							
2. Focus (Critical Thinking)	.61	-						
3. Evidence (Critical Thinking)	.56	.55	-					
4. Organization (Basics)	.54	.60	.63	-				
5. Organization (Critical Thinking)	.48	.50	.44	.48	-			
6. Style (Basics)	.48	.47	.53	.51	.40	-		
7. Style (Critical Thinking)	.47	.54	.51	.58	.44	.67	-	
8. Format (Basics)	.45	.38	.45	.42	.41	.43	.42	-

p<.001 for all correlations

If the traits behaved independently from one another, so that each of them could vary without affecting the others, we would see a very different pattern from the correlations in table 3. Namely, all the entries would be zero (the dashes indicate 1s, to indicate that a trait is trivially correlated with itself).

There are two possibilities to explain the fact that the correlations are not close to zero. One is that writers tend to produce papers such that if they are good in one area, they are good in another area, e.g. organization and style. The other explanation is that the peer reviewers have trouble distinguishing between the categories they are to rate and conflate them. There is no way to know with the data on hand which of these is the case, but it is probably a combination of the two explanations.

The next level of analysis is to delve within the correlation table and look for groups of traits that clump together within the ratings. For example, looking at table 3 again, we notice that all of the correlations are positive—

they all tend to go up together or down together. As an analogy, if we were to measure the height, length, weight, and food consumption of our pets, we would probably find that all these dimensions go up or down together, corresponding to the overall size of the pet. It turns out that the eight rubric ratings are predominately driven by a kind of size, which we take to be a holistic quality of the paper (see table 4).

The columns of table 4 give information for both peer ratings and instructor ratings. The most important figures are the component 1 columns. The top row shows that over half the variance in the eight individual trait scores is captured with a single number. Notice that the decimal numbers in each of the component 1 columns are nearly identical—all of them are around .3 to .4. This means that the main size component of the ratings is almost exactly a simple average of the eight individual scores. Interestingly, this is also how the grade is calculated for the paper.

There is little difference in this tendency between peers and instructors. It stretches the imagination to think that in the writing assignments, each of these eight traits would naturally progress at the same rate. We would expect some differences in development; maybe organization develops faster than style, for example. This tilts the explanation toward the other possibility, that raters are influenced by a holistic or average sense of the quality of the work and assign the eight individual ratings informed by that impression. If this is the case, then the rubric probably doesn't really need eight traits. Evaluating eight rubric criteria 46,689 times—that is, making 373,512 evaluations—represents a great deal of effort.

Ultimately, then, this finding challenges the notion that holistic scoring is invariably less helpful and precise than analytical rubrics. Clearly, there are strengths and weaknesses to analytic-trait scoring and holistic scoring: With the analytic approach, instructors may intuit the overall grade and work backwards to fill in rubric scores. This gives them a numerical explanation for grades assigned and, in theory, tells the student what he or she needs to improve. By contrast, the holistic/contrast approach would sacrifice this bookkeeping justification to some extent but also free up the instructor to give a wider range of feedback. It also makes measuring and training inter-rater reliability easier because there is only one holistic score to agree on, and more freedom can be granted to the contrasting strengths and weaknesses, as well as creative rule-breaking.

Table 4
Principle Components Analysis of Rubric Traits, First Two Dimensions

	Peer Ratings (n=30,377)		Instructor Ratings (n= 16,312)	
	Component 1	Component 2	Component 1	Component 2
Proportion of Variance	56%	9%	58%	11%
Focus (Basics)	.36	.30	.34	.40
Focus (Critical Thinking)	.37	.21	.38	.22
Evidence (Critical Thinking)	.37	-	.36	.26
Organization (Basics)	.38	-	.39	-
Organization (Critical Thinking)	.32	.48	.39	-
Style (Basics)	.35	-.58	.33	-.61
Style (Critical Thinking)	.37	-.51	.35	-.56
Format (Basics)	.30	.19	.28	.18

Discussion

The results provide mixed support for our peer-review practices in ENC1101 and ENC1102. On the one hand, it is encouraging to observe that students' reviews were more positively correlated with instructors' reviews over time, suggesting that either students and instructors are getting more adept at identifying quality writing or that students are getting a better sense of their instructor's preferences. These results echo the results of Liang and Tsai who found correlations between instructors and students improved over time on a four-point analytic scale, which suggests, not surprisingly perhaps, that practice improves peer reviews. Additionally, the finding that stronger writers, as identified by their instructors' grades on their papers, have scores more highly correlated with instructors and a broader variance in their scores than weaker students affirms peer-review pedagogy for more accomplished writers.

On the other hand, the difficulties students have distinguishing between B and C papers and the lack of variation in their ratings suggest there are problems with our peer-review practices. While the discrepancy

between student and instructor scores, particularly those reviews conducted by students who do not receive high grades for writing assignments, suggests instructors and students need to be skeptical of students' numeric evaluations of other students' work, the finding that peer reviews of students become more highly correlated with instructors over time suggests that there is some value to this practice. In addition, just because some students, particularly students who receive lower grades on their writing from instructors, may not be able to score like their instructors doesn't mean they aren't providing useful critical feedback or that going through this process isn't helpful in terms of helping them better understand their instructor's grading criteria or academic conventions for writing well.

Before assuming the lack of agreements between instructors and students invariably undermines the validity of peer review in general, we need to research the lexical comments peers provide one another. It could be that social pressures warrant inflated grades, yet the sticky notes, rubric dialog boxes, community comments, and endnotes may provide more critical, useful feedback. To research this point in the future, we are currently text mining students' comments and instructors from a lexical perspective. We are also working with colleagues at Malmö University in Sweden and University of Tartu in Estonia to look at cultural differences in peer reviews. Beyond conducting a lexical analysis of comments offered by peer reviewers and instructors at different university writing programs, we believe we need to measure the effects of comments and scores on revision before determining whether or not a discuss rubric is invariably superior to a numeric rubric. Alternatively, given that we found students and instructors may fixate on an overall value of a work being rated and then apply that holistic score to whatever sub-criteria are defined by a rubric, it could be that the psychology of assigning numbers in a column in this way creates a tendency to the mean because of an anchoring effect (Englich, Mussweiler, and Strack). If so, this argues that we should be more thoughtful about how rubric scales are constructed and what they are supposed to measure. Since conducting this study, we have modified My Reviewers to allow for more variation in rubrics and rubric scales.

Our results differ from those of Cho and Schunn, as well as Cho, Chung, King, and Schunn, who found that students found peer reviews superior to instructor reviews when at least six students conducted the reviews. Our results in this study, our analysis of teacher commentary (Dixon and Moxley), and our analysis of 52,001 essays scored by instructors in our program (Tackitt, Moxley, and Eubanks) do not support Cho et al.'s argument that instructors' expert status prevents them from providing the detailed, contextualized feedback students need and that they are likely to

underestimate the difficulty of revising. Perhaps this is due to a difference in context: We were working with English faculty while Cho et al. were working with STEM faculty. While possible, these disciplinary distinctions may not explain our divergent findings, so we would like to pursue this question in the future: thanks to USF funding, we are now examining peer review in STEM courses at USF, University of Pennsylvania, MIT, Dartmouth, and NCSU. Once this corpus develops, we will compare students' peer review experiences in STEM courses as well as English courses.

Because of this analysis of 46,689 student and instructor reviews of intermediate drafts and those of a related study that analyzes 52,001 scores provided by instructors on intermediate and final drafts (Tackitt et al.), we decided it was time to change the rubrics we use in first-year composition. As discussed above, the finding that a sizable percentage of any click was a holistic score suggested to us that we were asking our instructors and students to click too many criteria. While we stand by our earlier accounts regarding the surprising benefits of using a community rubric across genres and sections of ENC1101 and ENC1102 (see Moxley "Big Data"), beginning in the fall 2015 semester we have implemented genre-specific rubrics for our three projects in both courses. In the newest iteration of My Reviewers, we have accounted for all possible permutations: Administrators may now standardize rubric(s) across a program; alternatively, instructors may create rubrics with unlimited criteria. Administrators and instructors may now customize the scoring scale, dynamically adding as many milestones as they wish, with a minimum of two points and a maximum of 100 points (see fig. 5), and they can click on any part of the sliding scale to make more nuanced scoring determinations.

Despite this and related research, we remain somewhat conflicted regarding best peer review and writing program assessment practices. Not surprisingly, perhaps, we find ourselves oscillating between two dominant approaches to assessment: a modernist view that reifies grading and a postmodern view that embraces subjective responses to students' works. Of course, we are well aware of limitations with the modernist view, which assumes that descriptions of levels of achievement, combined with a system for collecting data, will produce ratings that correspond to the reified categories. Clearly, for example, the assumption that students (or instructors) will weigh the use of evidence in a piece of writing and produce a rating that corresponds to that (reified) construct overlooks legitimate differences in kinds and degrees of evidence needed by different readers or audiences. We understand that standardized assessments that are based in inauthentic,

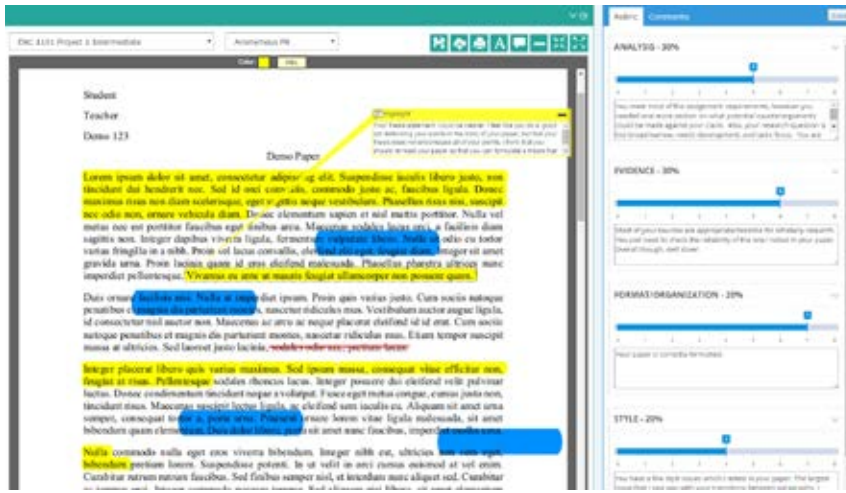

Fig. 5. Sample discuss rubric template.

out-of-class assignments and that strip texts from contexts may result in disingenuous, harmful claims. We understand that contemporary assessment practices are invariably racist and that students in the bottom quartile of the class struggle to improve against a headwind of negative feedback. We understand traditional rubrics have tended to focus on cognitive measures as opposed to addressing intrapersonal and interpersonal competencies.

Ultimately, however, society keeps score. Teachers, college admissions officers, accrediting bodies, legislators, employers, and governments all keep score. Instructors are required to provide grades. We know, for example, that the College Board determined in 2013 that 57% of SAT takers do not qualify as college ready ("2013 SAT Report" 3); that the ACT found 31% of high school graduates "did not meet any of the ACT College Readiness Benchmarks" ("The Condition of College" 4); that the NAEP Writing Report found 73% of 12th graders received scores of Below Basic or Basic as opposed to Proficient or Advanced in 2011 (National Center for Education Statistics); that the US literacy rate fell from 10th to 20th in the latest study on global rankings, Programme for International Student Assessment 2012 (PISA 2012). Clearly, these assessments have economic consequences. When it comes to employment, it matters that Blacks and Hispanics score significantly lower than whites on tests such as the NAEP assessment of writing. Moving forward, we are committed to further exploring how we can develop Writing Analytics at My Reviewers to identify data patterns for administrators, instructors, and students that can be used to improve peer review practices. We question how Writing Analytics, which repur-

pose data in the system (user trace patterns, lexical comments, scores, survey responses) can help writers and reviewers navigate the middle ground between extreme subjective or objective views of assessment. One possible approach is suggested by so-called non-cognitive measures, where we do not solely focus on isolated products of student work as proxies for some platonic student ability, but treat attitudes, beliefs, and behaviors related to writing as equally important. This greatly broadens the potential for meaningful assessment and pedagogy that impacts students in new ways.

Future Research

We believe a richer portrait of peer-review processes needs to be developed, one that accounts for students' and instructors' written comments and the effects of these comments on subsequent drafts, and one that accounts for lexical comments as associated with particular user attitudes, beliefs, and behaviors. We need lexical, descriptive work on what instructors and students are saying to one another and how these comments and scores support writing development. After all, students' comments on one another's papers may have a profound effect on the development of students' intrapersonal, interpersonal, and cognitive competencies. While a grade provides students information regarding the quality of their work in relation to their peers, students need thoughtful in-text and endnote comments to improve as writers.

Beyond employing sentiment analysis of the lexical corpus, future studies need to examine if a large set of students and instructors from different schools and different backgrounds will provide the same or similar results. We believe it would be especially helpful to evaluate how particular cohorts perform, such as students with high GPA or AP English scores or students from an L2 background. We also think it would be useful to study the response styles of instructor groups by disciplinary lens, such as literature versus creative writing versus Professional and Technical Writing versus Rhetoric and Composition. Perhaps the permissions at My Reviewers should permit students to see other students' scoring and markup as well as the instructors' scoring and markup—absent the instructor's grades, given FERPA—after the instructor grades peer review efforts.

Furthermore, we need experimental work that develops and tests algorithms and workflows. As symbol analysts, tool developers, and writing program administrators embedded in writing programs, we wonder what information digital tools such as My Reviewers can provide to facilitate better reviews, writing development, and transfer of cognitive, intrapersonal, and interpersonal competencies.

WPA: Writing Program Administration, Volume 39, Number 2, Spring 2016

Notes

1. Professor Moxley wishes to disclose a potential conflict of interest. While the My Reviewers software is not commercially available, it may become commercially available in the future. Because the data collection methods used in this study demonstrate the viability of My Reviewers, this research study may enhance the commercial value of My Reviewers. Ultimately, USF owns My Reviewers; however, Moxley possesses the rights to license My Reviewers. Professor Moxley has filed the necessary USF conflict of interest paperwork. The Conflict of Interest Committee at USF has developed a management plan with which Dr. Moxley has complied prior to submitting this and similar research.

2. Following the 2009–2010 academic year, 10 independent scorers reviewed the third/final drafts of 249 students' essays in ENC1101 and 249 essays in ENC1102. In other words, the same 249 students were tracked for a year, and their essays for Project 2 in ENC1101 were compared with their essays for Project 2 in ENC1102. The independent evaluators were unaware of the students' identities, unaware of the students' instructors' identities, and unaware of the scores provided by the students' classroom instructors. A comparison between the two external evaluators and the students' classroom instructor revealed few differences between the classroom instructors and independent scorers on 7 of the 8 rubric measures. The only discrepancy between the instructors and the outside evaluators was the Style/Basics subcategory: On that measure, the students' classroom instructors were tougher in their judgments—about a third of a grade tougher.

3. It is interesting to note that in *Agency in the Age of Peer Production*, the qualitative study that historizes our effort to crowdsource our curriculum from a datagogical lens, we chronicle instructor resistance to a generic rubric. Now, when we introduced the idea of changing the generic rubric, for making rubrics distinct for each project, we experienced resistance. USF had been praised by the Southern Association of Colleges of Schools Commission on Colleges during its accreditation review, and the Office of Institutional Effectiveness, which had partnered with us in the crowdsourcing effort to develop the rubric, was happy with the rubric. After numerous years of using one rubric to assess across genres and sections of ENC1101 and ENC1102, instructors had grown accustomed to and comfortable with our rubric. When we initially suggested changing the rubric again, we sensed that people hoped we would leave well enough alone. To us, this suggests instructors derive benefits from a stable curriculum over time.

4. When instructors establish their peer review groups, they can choose between a discuss version and a numeric version of the rubric. The discuss version provides dialog spaces with grades, and the numeric version provides dialog boxes and grades. The default view is the numeric version.

5. Excluding book collections and conference proceedings, journals include *Assessment & Evaluation in Higher Education*, *British Educational Research Journal*, *Assessing Writing*, *Turkish Online Journal of Educational Technology*, *Issues in Education Research*, *Active Learning in Higher Education*, *International Journal for the*

Scholarship of Learning and Teaching, *Journal of Writing Research*, *British Journal of Educational Technology*, *British Educational Research Journal*, *Instructional Science*, *The Internet and Higher Education*, *Journal of Educational Technology and Society*, *Journal of Second Language Writing*, *Computers and Education*, *Studies in Higher Education*, *Language Teaching Research*, *Journal of Second Language Writing*, *Review of Educational Research*, and *Internet and Higher Education*.

6. Thanks to the recent development of corpus-based methods to record and analyze students' assessments, new methodologies now enable researchers to analyze large-scale studies of students' evaluations and teachers' comments. Rather than base results on small sample sizes, typically 5% of the population, the entire population can be researched. The behaviors of the population in a digital environment can also be passively recorded and researched within the limits of user agreements. Given all data is provided in real-time, digital tools shatter the traditional bifurcation of formative from summative evaluations.

7. The Community Comments are extensive, and we provide them in a book-length etext format as well as in the database format. Each Community Comment is a clickable hyperlink that leads to a resource page that defines the comment in an article and video and then gives students an opportunity to test their comprehension via an online quiz. A typical resource page includes the following sections pertaining to the topic: definition, identification, revision, common types, usage conventions, videos, activities, and external links. Most of the existing Community Comments address composition matters, but we are working with STEM faculty across the disciplines at USF, NCSU, MIT, Dartmouth, and Penn to build comments for other communities.

8. During the time of this study, the instructors who graded peer reviews assigned A grades 85% percent of the time. Many instructors ignored the policy requirement to grade peer reviews. The writing program administrators chose not to strictly enforce the policy that peer reviews should be graded by instructors because they have chosen to adopt a soft power approach as outlined in *Agency in the Age of Peer Production*.

9. At the time of this study, we did not have IRB approval to match demographic data to user behaviors. We have since received that permission and are working on additional studies that maps behaviors by demographics and other variables, including grit, self-efficacy, and self-regulation.

Works Cited

ACT. "The Condition of College & Career Readiness 2013." *ACT* (2013): 1–32. Web. 8 Feb. 2016.

Anson, Chris M., Deanna P. Dannels, Pamela Flash, and Amy L. Housley Gaffney. "Big Rubrics and Weird Genres: The Futility of Using Generic Assessment Tools Across Diverse Instructional Contexts." *Journal of Writing Assessment* 5.1 (2012): n. pag. Web. 13 Mar. 2016.

Anson, Chris, Joseph Moxley, Djuddah Leijen, Damian Finnegan, Anna Wärnsby, and Asko Kauppinen. "Theorizing Community Rubrics: Limits, Research, and Case Studies." *8th Biennial Conference of the European Association for the Teaching of Academic Writing*. Tallinn University of Technology, Estonia. 15–17 June 2015. Presentation.

Boase-Jelinek, Daniel, Jenni Parker, and Jan Herrington. "Student Reflection and Learning through Peer Reviews." *Teaching and Learning in Higher Education: Western Australia's TL Forum*. Spec. issue of *Issues in Educational Research* 23.2 (2013): 119–31. Print.

Cho, Kwangsu, Tingting Rachel Chung, William R. King, and Christian Schunn. "Peer-Based Computer-Supported Knowledge Refinement: An Empirical Investigation." *Communications of the ACM* 51.3 (2008): 83–88. Print.

Cho, Kwangsu, and Christian D. Schunn. "Scaffolded Writing and Rewriting in the Discipline: A Web-Based Reciprocal Peer Review System." *Computers & Education* 48.3 (2005): 409–26. Print.

College Board. "2013 SAT Report on College and Career Readiness." *College Board* (2013): 1–9. Web. 8 Feb. 2016.

Crossman, Joanne M., and Stacey L. Kite. "Facilitating Improved Writing among Students through Directed Peer Review." *Active Learning in Higher Education* 13.3 (2012): 219–29. Print.

Dixon, Zachary, and Joseph Moxley. "Everything Is Illuminated: What Big Data Can Tell Us about Teacher Commentary." *Assessing Writing* 18.4 (2013): 241–56. Web. 13 Mar. 2016.

Englich, Birte, Thomas Mussweiler, and Fritz Strack. "Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making." *Personality and Social Psychology Bulletin* 32.2 (2006): 188–200. Print.

Esfandiari, Rajab, and Carol M. Myford. "Severity Differences Among Self-assessors, Peer-assessors, and Teacher Assessors Rating EFL Essays." *Assessing Writing* 8.2 (2013): 111–31. Print.

Eubanks, David. "A Geometric Approach to Conditional Inter-Rater Agreement." 2015. TS. Furman University, South Carolina.

Falchikov, Nancy, and David Boud. "Student Self-Assessment in Higher Education: A Meta-Analysis." *Review of Educational Research* 59.4 (1989): 395–430. Print.

Falchikov, Nancy, and Douglas Magin. "Detecting Gender Bias in Peer Marking of Students' Group Process Work." *Assessment & Evaluation in Higher Education* 22.4 (1997): 385–96. Print.

Falchikov, Nancy, and Judy Goldfinch. "Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks." *Review of Educational Research* 70.3 (2000): 287–322. Print.

Haswell, Richard H. "NCTE/CCCC's Recent War on Scholarship." *Written Communication* 22.2 (2005): 198–223. Print.

Khonbi, Zainab Abolfazli, and Karim Sadeghi. "The Effect of Assessment Type (Self Vs. Peer) on Iranian University EFL Students' Course Achievement." *Procedia-Social and Behavioral Sciences* 70 (2013): 1552–64. Print.

Langbehn, Karen, Megan McIntyre, and Joe M. Moxley. "Re-Mediating Writing Program Assessment." *Digital Writing Assessment & Evaluation*. Ed. Heidi McKee and Dànielle Nicole DeVoss. Logan: Computers and Composition Digital Press, n. pag. Web. 13 Mar. 2016.

Li, Lan, Liu Xiongyi, and Yuchun Zhou. "Give and Take: A Re-Analysis of Assessor and Assessee's Roles in Technology-Facilitated Peer Assessment." *British Journal of Educational Technology* 43.3 (2012): 376–84. Web. 13 Feb. 2016.

Liang, Jyh-Chong, and Chin-Chung Tsai. "Learning through Science Writing via Online Peer Assessment in a College Biology Course." *The Internet and Higher Education* 13.4 (2010): 242–47. Print.

Liu, Eric Zhi-Feng, and Chun-Yi Lee. "Using Peer Feedback to Improve Learning via Online Peer Assessment." *Turkish Online Journal of Educational Technology* 12.1 (2013): 187–99. Print.

Liu, Xiongyi, and Lan Li. "Assessment Training Effects on Student Assessment Skills and Task Performance in a Technology-Facilitated Peer Assessment." *Assessment & Evaluation in Higher Education* 39.3 (2014): 275–92. Print.

Lu, Yingjie, Pengzhu Zhang, Jingfang Liu, Jia Li, and Shasha Deng. "Health-Related Hot Topic Detection in Online Communities Using Text Clustering." *PLoS ONE* 8.2 (2013): n. pag. Web. 13 Feb. 2016.

Lundstrom, Kristi, and Wendy Baker. "To Give Is Better Than to Receive: The Benefits of Peer Review to the Reviewer's Own Writing." *Journal of Second Language Writing* 18.1 (2009): 30–43. Print.

Moxley, Joseph M. "Aggregated Assessment and 'Objectivity 2.0.'" *Proceedings of the Second Workshop on Computational Linguistics and Writing (CLW 2012): Linguistic and Cognitive Aspects of Document Creation and Document Engineering*. 23 Apr. 2012: 19–26. Web. 13 March 2016.

—. "Big Data, Learning Analytics, and Social Assessment." *Journal of Writing Assessment* 6.1 (2013): n. pag. Web. 13 Feb. 2016.

National Center for Education Statistics. "The Nation's Report Card: Writing 2011." *National Center for Education Statistics*. Institute of Education Sciences, US Department of Education, Washington, DC. Sept. 2012. Web. 8 Feb. 2016.

Patchan, Melissa M., Christian D. Schunn, and Russell J. Clark. "Writing in Natural Sciences: Understanding the Effects of Different Types of Reviewers on the Writing Process." *Journal of Writing Research* 2.3 (2011): 365–93. Print.

Programme for International Student Assessment. "Programme for International Student Assessment (PISA) Results from PISA 2012: United States." *OECD*. PISA, n.d. Web. 8 Feb. 2016.

Schneider, Barbara, Martin Carnoy, Jeremy Kilpatrick, William H. Schmidt, and Richard J. Shavelson. *Estimating Causal Effects: Using Experimental and Observational Designs*. Washington, DC: American Educational Research Association, 2007. Print.

Tackitt, Alaina, Joe M. Moxley, and David Eubanks. "Big Rubrics and Big Data: The Development, Transfer, Prediction, and Restriction of Student Competencies." 2015. TS. *Assessing Writing*.

Takeda, Sachiko, and Fabian Homberg. "The Effects of Gender on Group Work Process and Achievement: An Analysis through Self- and Peer-Assessment." *British Educational Research Journal* 40.2 (2014): 373–96. Print.

Topping, Keith J. "Peers as a Source of Formative and Summative Assessment." *SAGE Handbook of Research on Classroom Assessment.* Ed. James H. McMillan. Thousand Oaks: SAGE, 2012. Print.

—. "Peer Assessment between Students in Colleges and Universities." *Review of Educational Research* 68.3 (1998): 249–76. Print.

Topping, Keith J., Elaine F. Smith, Ian Swanson, and Audrey Elliot. "Formative Peer Assessment of Academic Writing between Postgraduate Students." *Assessment and Evaluation in Higher Education* 25.2 (2000): 149–69. Print.

Tucker, Richard. "Sex Does Not Matter: Gender Bias and Gender Differences in Peer Assessments of Contributions to Group Work." *Assessment & Evaluation in Higher Education* 39.3 (2014): 293–309. Print.

Vierrege, Quentin D., Kyle D. Stedman, Taylor Joy Mitchell, and Joseph M. Moxley. *Agency in the Age of Peer Production*. Urbana: NCTE, 2012. Print.

Acknowledgments

We thank Norbert Elliot for his critical feedback and encouragement on this article. For their vigorous editorial work article on this article, we thank Barbara L'Eplattenier, Lisa Mastrangelo, Sherry Rankins-Robertson, and Davee Sarim. Finally, we thank Val Ross who has played a critical role in the on-going development of My Reviewers. This research is supported by the National Science Foundation under Award #1544239, "Collaborative Research: The Role of Instructor and Peer Feedback in Improving the Cognitive, Interpersonal, and Intrapersonal Competencies of Student Writers in STEM Courses."

Joseph M. Moxley <http://joemoxley.org> is the founder of Writing Commons *<http://writingcommons.org>, a free alternative to expensive writing textbooks. Peer-reviewed,* Writing Commons *provides open access to over 1,000 webtexts, making it a viable choice as the required textbook for composition, professional and technical writing, creative nonfiction, and creative writing courses. Moxley is also Director of First-Year Composition at the University of South Florida, a Research 1 university.*

David Eubanks is Assistant Vice President for Assessment and Institutional Effectiveness at Furman University, a private liberal arts university.