Signifying Scores: Instructor Rating as an Assessment Measure

Alaina Tackitt (Eckerd College),
Joseph M. Moxley (University of South Florida),
and David Eubanks (Furman University)

Corresponding author: Alaina Tackitt
University of South Florida
Department of English, Cooper Hall 107
4202 East Fowler Avenue
Tampa, FL 33620
(813)-367-6020
adtackitt@gmail.com

Alaina Tackitt is the Director of Writing Services for Eckerd College, Program for Experienced Learners and a candidate for the Ph.D. in English, Rhetoric and Composition at the University of South Florida; she is currently writing her dissertation on adult learners in Composition.

Joe Moxley <http://joemoxley.org> is the Founder of *Writing Commons* <http://writingcommons.org>, an open access alternative to expensive textbooks. Peer-reviewed, *Writing Commons* provides over a thousand webtexts, making it a viable 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.

David Eubanks is Assistant Vice President for Assessment and Institutional Effectiveness at Furman University. Eubanks develops predictive analytics and ways to assess student achievement authentically.

Abstract
The place of digitally-evaluated, rubric-assisted instructor scoring of student writing has not been situated within conversations questioning the use of instructor evaluations as instruments of programmatic assessment. This study gauges the viability of utilizing such scores for programmatic assessment by analyzing classroom-based scores generated by 128 instructors on 52,001 intermediate and final essays written by 7,722 students spanning 3 years, 7 terms, and 482 sections to test the validity of the rubric construction and the reliability of the instructor evaluations. The analysis investigates subscores and their relation to average rubric ratings, examines the correlation between subscores, and explores instructor scoring over two terms. The findings challenge the construct validity of the rubric and suggest that the subscores are being conflated, demonstrate that both instructor grading styles and student traits are somewhat persistent across time and that scores are predictive, and reveal the reset effect across terms. Locating and surveying the intersection of the rubric construction and instructor scoring contributes to the larger debate surrounding instructor evaluations and assessment by clarifying the application of rubrics, complicating the validity of instructor scoring, and advancing questions surrounding the use of instructor scores for programmatic assessment.

Keywords: Writing Assessment, Programmatic Assessment, Instructor Scoring, Rubrics

1. Introduction

Digital tools that enable instructors to grade and comment on student papers and peer reviews online are transforming how instructors and students critique documents and have the potential to transform how writing and writing programs are assessed. Beyond profoundly altering how faculty and students respond to writing, these tools aggregate e-portfolios, facilitate distributive evaluation, and archive data that allow researchers to mine texts and map student outcomes in order to produce analytics that inform users, researchers, and administrators. While the initial purpose of such tools is the evaluation of student work and the immediate product is feedback generated to culminate in a course grade, it is possible to utilize the resulting scores for programmatic assessments.

Programmatic assessment, broadly, is assessment beyond, and often of, the classroom; such assessment of learning for learning can include the assessment of students, instructors, course material and structure, course tools, or other areas of interest and can be performed by program administrators, university administrators, outside accrediting bodies, or other stakeholders. Programmatic assessment can be unrelated to course evaluations or can include instructor evaluations. Research on programmatic assessment both champions (Condon & Kelly-Riley, 2004; Mason & Drag, 2010) and challenges (Carter, 2003; Elliot, 2005; Rogers, 2003) the use of instructor evaluation—specifically, the evaluations of student writing—for programmatic assessment.

Data captured by digital tools offer a lens through which to gauge the application of classroom-based, rubric-assisted instructor evaluation of writing for programmatic assessments. Investigators have researched the reliability of instructor evaluation of writing (Gentile, 2000; Sax, 1980; Sommers, 1982; Starch and Elliot, 1912, 1913a, 1913b) and the utilization of rubrics for the evaluation of writing (Anson, Dannels, Flash, & Housley, 2012; Elbow, 2006; Inoue, 2014; Moxley, 2013), but such conversations and questions have not converged in connection to the assessment applications of digital tools such as rubric-assisted instructor evaluation. In order to test the viability of employing instructors' scores, specifically scores generated digitally using a standard rubric, as means or programmatic assessment, this investigation analyzes a data set of 52,001 essays written by 7,722 students spanning 2 courses over 3 years, 7 terms, and 482 sections of first-year composition in order to deconstruct and interpret the rubric scores of 128 instructors. Examining the construct validity and reliability of instructor-assigned rubric scores reveals the relation between using scores as course grades and for programmatic assessment. While the results clarify the use of rubrics, they complicate the validity of instructor scoring and refine considerations regarding the use of instructor scores for programmatic assessment.

2. Review

The evaluation of student learning through student writing is a modern model made possible through modern means and methods. Until nearly the 20th century, examination was mainly oral (Thayer, 1928). In the 5th century B.C.E., reciting Euripides earned Athenian prisoners freedom from quarries (Plutarch, 29.2). Writing, meanwhile, was dismissed as a childish exercise to facilitate memorization, or worse, as Plato warned, deemed dangerous (Swearingen, 1991). By the 5th century C.E., classical recitation was required for introductory-level Chinese civil service positions (Cressey, 1929). When writing was taught, the focus was primarily on the development of penmanship (Thayer, 1928). The modern practices of teaching

composition and evaluating student writing are rooted in the development of technologies that expanded educational options and opportunities, created and divided disciplines, and ignited debates regarding instruction, evaluation, and assessment.

With the invention of the metal pen, movements toward written examinations were prompted, and a variety of approaches to evaluation, including numerical scoring, became an option. Disciplines were divided on the theory and practice of evaluation and grading. Mathematical examinations, for instance, included *right* answers, which made earning a maximum score possible and evaluation clear. When evaluating Literature and History, on the other hand, earning the maximum score was considered unlikely, and it was suggested that 10% of the grade be reserved for the more subjective and sensitive matters of style and finesse (Cureton, 1971). Subjects considered subjective were seen as complicated, perhaps sophisticated, but the subjective was not seen as suspect.

The developing practices of evaluating and grading written work led to new methods and processes in an attempt to systemize the praxis of classroom-based student evaluation; they also led to new practices of examining teachers and assessing programs and schools. In an early example, a committee surveying the Boston schools not only made claims of cheating on examinations but also levied more serious charges of guessing. While some scholars and students were conscientious enough not to answer in ignorance, others recklessly concocted responses hoping to earn some credit and appear informed (Caldwell and Courtis, 1971). Although accusations of guessing as intellectual fraud have faded, arguments surrounding evaluation and assessment flourish, specifically in relation to the evaluation and assessment of written work. Complicating considerations of classroom evaluation and programmatic assessment are concerns of conflating grading and scoring. Some academics and administrators suggest that programmatic assessment should not stem from classroom evaluations, while others argue that instructor evaluation can serve to review both student progress and programmatic success.

Specifically, researchers in Engineering offer evidence in support of using instructor scoring of student coursework not only for classroom and programmatic evaluation but even for externally mandated programmatic assessment. Mason and Dragovich (2010) develop and test a "coursework based assessment methodology" to be used both for programmatic assessment and to fulfill the Accreditation Board for Engineering and Technology's (ABET) Engineering Accreditation Commission (EAC) assessment requirement; their finding suggest that given appropriate analysis, "grades based on a student's coursework, such as exams, homework, labs, and papers, can be useful" assessment measurements (p. 206). Several advantages to using instructor evaluations for programmatic assessment emerge, such as limiting overhead and offering direct evidence on potential programmatic improvements while providing continuous data (p. 207). In general, course-embedded assessments and standardized rubrics have also proved successful as means of assessing General Education Programs (Gerretson & Golson, 2005). There are, however, significant theoretical and practical concerns with employing classroom evaluations for programmatic assessment.

Much of the evidence against utilizing course grades for programmatic assessment contends that students' scores are not a reliable means of assessing programs because classroom objectives often differ from programmatic goals and grading policies may vary from course to course (Carter, 2003). While it is acceptable that the formative assessment drawn from classroom evaluations serve as a factor in the overall programmatic assessment, classroom evaluations should not be the whole of the assessment, and instructors should not be the only evaluators of programmatic progress (Rogers, 2003). Ultimately, critics argue that assessment should be big

picture and long term and transfer beyond the classroom. In effect, despite the fact that both students and institutions place great importance on instructors' scores as evaluative of student success, some academicians, administrators, and accreditors minimize and marginalize the value of instructor evaluation. Contextual analysis suggests that assumptions and assertions supporting claims that instructors' evaluations are unacceptable measures of programmatic assessment are grounded in a general suspicion of instructor evaluation.

Willingham, Pollack, and Lewis (2002), research scientists from ETS, provide an extensive review of literature demonstrating long-held distrust of instructor feedback and scoring. Similarly, in their review of the history of writing assessment, Huot, O'Neill, and Moore (2010) claim that "one hundred years ago, teacher judgments were found suspect, and a test was assumed to be better at helping university admissions personnel make important, consequential decisions about students" (p. 497). Institutional and departmental accreditations play a role in defining the value of instructor evaluation in the classroom and in relation to programmatic assessment, and the disciplinary divide evident in early conversations about grading has persisted and perhaps expanded. Contrary to the example from Engineering, where the ABET accepts the classroom evaluations for accreditation (including the evaluation of student papers), the distrust of instructor evaluation in Writing Studies is grounded not only in a general distrust of teachers and their evaluations but in a specific suspicion of the evaluation of student writing and of the evaluators of writing.

Starch and Elliot's (1912, 1913a, 1913b) seminal works on the unreliability of instructor evaluation of essays have been replicated since publication (Sax, 1980), and more recent research confirms that feedback is subjective even when a rubric is applied and suggests that evaluation should be left to subject-matter experts instead of writing instructors (Gentile, 2000). Research within Writing Studies also devalues the evaluation of writing instructors. In her landmark essay on instructor responses to writing, Sommers (1982) concludes that most instructor feedback amounts to vague commenting absent useful, text-specific critiques. Such dismissal of teachers and teacher evaluations reinforces the arguments against classroom-based evaluations as measures of assessment and promotes assessments that are created and evaluated far from the teachers and classroom. Elliot (2005) traces the absence of teacher authority in traditional assessment methods to the establishment of the CEEB and links it to the policies of the College Board that devalue teachers' grades and emphasize SAT scores. The disciplinary divide evident in early conversations and considerations in relation to grading exist today in the continued tension between standardization and subjectivity. Attempts at standardization have moved into the classroom, as well, often in the form of rubrics.

The debates and distrust surrounding rubrics mirror discussions and allegations regarding instructor evaluations. Researchers in Writing Studies are openly suspicious of rubrics as a more objective means to measure the development of student writing and thinking (Inoue, 2014; Turley & Gallagher, 2008). Efforts to standardize responses have been classified as pernicious, undermining what is inevitably a subjective activity (Elbow, 2006; Wilson, 2007), and research has implied that raters often ignore rubrics and score in response to the preponderance and frequency of error (Rezaei & Lovorn, 2010). Especially controversial is the use of a singular rubric deployed across genres and disciplines (Anson, Dannels, Flash, & Housley, 2012; Moxley, 2013). Arguments range from calls for the "education-wide abandonent of generic rubrics" to suggestions for "contextually-based approaches to assessment" (Anson, Dannels, Flash, & Housley, 2012, para. 3) and recommendations that students design their own rubrics for writing assignments (Inoue, 2005). Given the ecosystem in which instructor evaluations and

rubrics are distrusted, contemplating the use of rubric-assisted instructor ratings for programmatic assessment feels unwarranted, but placing the data within the disciplinary conversations regarding assessment exposes the great potential of using the instructor scores under consideration and by extension, scores constructed similarly.

Writing Studies recognizes that objective tests are not successful in evaluating students' ability to write, obviating the function of standardized testing, but the struggle to reconcile the subjective through standardization persists, and while portfolios are accepted as a means of programmatic assessment, they are considered "messy" as a result of the multiple texts and types of texts that make them hard to evaluate, especially given the constraints of collaborative or communal scoring (Yancey, 1999). The data set being assessed is functionally comprised of e-portfolios from both first-year composition courses that represent every student and instructor, which corrects for concerns by including classroom-based, contextual evaluation while simplifying the possibility for subsequent multiple or external evaluations. Additionally, representatives of the Conference on College Composition and Communication (CCCC) warn against indirect assessment (CCCC, 2014), and the National Council of Teachers of English and the Council of Writing Program Administrators caution that the assessment of writing must account for context (CWPA, 2008). The scores being analyzed have the potential to serve as assessment measures that fulfill disciplinary expectations for both classroom evaluation and programmatic assessment by accessing direct, classroom evaluation for programmatic assessment.

The CCCC Committee on Assessment, the National Council of Teachers of English, and the Council of Writing Program Administrators have all weighed in on the assessment of writing and how it can and should be done well. In alignment with conditions and provisions published by the CCCC Committee on Assessment in the "Writing Assessment: A Position Statement," the scores encompassed within the data set respond to writing across "a range of purposes for a range of audiences in a range of settings," represent the culmination of student writing that demonstrates the "sum of a variety of skills employed in a diversity of contexts" to balance fluctuations in ability, and reflect direct assessment by human readers using multiple measures including "assessment by peers, instructors, and the student" (CCCC, 2014). And as required in the first guiding principle, assessment through these scores has proved to improve teaching and learning (Langbehn, McIntyre, & Moxley, 2013). The digital capture of instructor-generated, rubric-assisted, classroom-based scores allow for assessment that embodies guiding principles and exemplifies many of best practices for writing assessment.

Similarly, in the "NCTE-WPA White Paper on Writing Assessment in Colleges and Universities," the National Council of Teachers of English and the Council of Writing Program Administrators present their "principles of effective writing assessment." After recognizing the highly contextual nature of assessment measures, this document offers the elements generally accepted by members of both organizations that create "effective, meaningful, and responsible writing assessment"; again, the set of scores under consideration offers a great opportunity to construct and conduct assessments that fulfill disciplinary expectations by providing "a foundation for data-driven, or evidence-based, decision making" that employs "multiple measures and engage[s] multiple perspectives to make decisions that improve teaching and learning,"—measurements that "encourage and expect teachers to be trusted, knowledgeable, and communicative" (CWPA, 2008). But for instructor scores to be widely accepted as a viable option for programmatic assessment in Writing Studies, the disciplinary distrust of rubrics and the generalized distrust of instructor evaluations must be addressed, if not alleviated

The practical debate surrounding the question of whether or not to use classroom evaluations beyond the classroom, when contextualized, appears to be a microcosm manifesting from a theoretical disagreement on the value of teachers and perhaps even the value of different disciplines and their methods and materials. The distrust of rubrics, of instructors, and of assessment—especially in Writing Studies—undergirds the distrust of using classroom evaluations for programmatic assessment. Questioning the use of teacher evaluations in programmatic assessment, then, requires addressing all of these considerations in concert by examining rubric construction and application in conjunction with instructor application of tools and the reliability of scoring in the context of using these rubric-based, classroom evaluations for programmatic evaluation.

Digital tools are changing the way student writing is evaluated, which challenges considerations surrounding the use of instructor evaluations of student writing for programmatic assessment and requires common questions to be reevaluated and current conversations to be expanded. Scholarly conversations research the use of instructor evaluation for programmatic assessment, but questions surrounding the use rubric-assisted instructor ratings for programmatic assessment have not been added to these conversations. The programmatic design and the application of a web-based suite of tools allow the data being analyzed to correct for a number of major arguments against using instructor scores for programmatic assessment through the standardization of grading policies and course material across simultaneous sections and through the creation of a comprehensive and varied cache of student writing—essentially producing an environment conducive to the creation of corrective conditions and providing an ideal scoring situation of study in relation to the usability of instructor scores for programmatic assessment. The results of analyzing the existing data set complicate discourses of distrust of surrounding the evaluation of writing and the use of rubrics and, ultimately, reveal the limits of using instructors' scores for programmatic assessment.

3. Methods

3.1 Setting

This investigation takes place within the First Year Composition (FYC) program of a large, public university in the United States. The primary coursework of the program consists of two courses, ENC1101 and ENC1102 (hereafter 1101 and 1102), which are designed as a sequence and include three common, major projects. Students write at least three drafts of each project and receive feedback at least three times from their instructors (http://hosted.usf.edu/fyc/). Additionally, each student meets with his or her instructor at least twice each semester to discuss these projects individually and conducts at least one round of peer review for each project using the community rubric. In general, the curriculum is standardized across the data set: students write three major papers and instructors respond to three drafts (one for each project). Although the three projects differ in detail somewhat from year to year, they are consistently designed to progress across both terms and function as six graduated assignments that increase in difficulty.

The projects in 1101 evolve from learning research skills to synthesizing the arguments of others and end with the deconstruction of an argument. Beginning where 1101 concludes, 1102 starts with the deconstruction of multiple arguments on a single topic and moves through the construction of an argument to the development of an argument including a call for an action.

The student learning outcomes driving the assignments in 1101 include critical reading, summarizing, synthesizing, and thesis development while1102 aims to advance the growth of the rhetorical skills of argument, voice, and persuasion in order to develop student agency. In each, students are challenged but not required to choose a topic they find interesting, to research that topic over the course of a semester, and to write about it in different genres. For all six projects, a single, community rubric is used for assessment and grading purposes. While 1102 can function independently (for students who do not take 1101), it is designed as a continuation of 1101.

3.2 Sample

Subjects in this study come from a population of university students enrolled in a composition courses (1101 or 1102) at USF in Tampa, Florida. According to IPEDS data, undergraduates at USF-Tampa have an average high school GPA of 3.78, median SAT reading scores of 570, and median SAT writing scores of 550. Approximately 38% of applicants are admitted, and 88% of first-year students return for the second year. Florida's statewide university policies allow new undergraduates to exempt 1101 or both 1101 and 1102 depending on equivalency requests, SAT scores, or AP English scores. Student demographics may vary slightly over the Fall, Spring, and Summer terms and from year to year. Data on these potential differentiators was not provided by the university, so the possible confounding effects of theses unknowns is a limitation of the present work and potential for future and further study.

Nearly all of the FYC instructors are graduate students working as teaching assistants who range from Master's-level students to Doctoral candidates. In addition, approximately five of the instructors each semester are adjuncts who teach between one and four sections. During the three-day orientation for all FYC instructors and a subsequent training for new instructors, workshops are held on responding to writing and using the rubric to score essays. All new instructors—approximately twenty-five each year—participate in a semester-long practicum. In addition, each new instructor is assigned a mentor who reviews the new instructors' comments on and scoring of student work and provides feedback. Writing program administrators also review instructors' scores and comments for mentoring purposes. Numerous supporting resources have been developed including extensive common comments, videos, marked up sample essays, and articles about rubric criteria.

3.3 Measures

All instructors use the same rubric to assess all papers (see Table 1). The rubric was developed via a crowd-sourcing process that involved discussions with instructors, program administrators, the Office of Institutional Effectiveness, and the General Education Council (Vieregge et al. 2012).

Table 1

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
	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

The rubric contains five core criteria: Focus, Evidence, Organization, Style, and Format. For three of these criteria, the rubric contains two subcategories: Basics and Critical Thinking. The categories Focus, Organization, and Style are internally divided into two subscores termed Basics and Critical Thinking. The Format category is categorized as Basics while the Evidence category is designated as Critical Thinking, which allows instructors to respond to a total of eight subcategories. As a result, the rubric comprises eight individual ratings of a student paper, which are placed into five categories: Focus (the two subscores R1 and R2), Evidence (R3), Organization (R4 and R5), Style (R6 and R7), and Format (R8) as shown in Table 2. The intent is to measure five aspects of student writing, and within three of these five, to distinguish between different levels of achievement. Consequentially, the Basics and Critical Thinking subscores could be assessed as two, distinct categories since the four subscores R2, R3, R5, R7 are categorized as representative of critical thinking and R1, R4, R6, R8 are designated as characterizing basic skills. The project grade is determined by an average of the eight subscores, which we denote Ravg.

Table 2

Data Description

Variable: Meaning		
Term: Year & term of class	Draft: Intermediate or Final	R5: Organization (critical thinking) (0-4)
Class: ENC1101, ENC1102	R1: Focus (basics) (0-4)	R6: Style (basics) (0-4)
StudID: Student identifier	R2: Focus (critical thinking) (0-4)	R7: Style (critical thinking)(0-4)
ProfID: Instructor identifier	R3: Evidence (critical thinking (0-4)	R8: Format (basics) (0-4)
Project: 1, 2, or 3	R4: Organization (basics) (0-4)	Ravg: Average of R1 through R8

3.4 Data collection

Students upload essays to My Reviewers, <http://myreviewers.com>, a digital tool developed at USF to facilitate document reviews, peer reviews, team projects, and portfolios. Document markup tools allow instructors to use the rubric to assess the primary coursework including intermediate and final drafts of six major projects spanning genres such as annotated bibliographies, literature reviews, analytic essays, historiographies, Rogerian arguments, remediations, and arguments for social justice. Records of instructor ratings are queried from the database, resulting in a sequence of rating *events*, each with its own row of data. An event

comprises the data elements indexed in Table 1. The total data set comprised 53,042 essay score sets. After removing problematic samples, what remained was 52,001 intermediate and final essays from 7,722 students spanning 7 terms and more than 3 years to total 482 course sections (the average class size is approximately 20) taught by 128 instructors.

3.5 Data analysis

As illustrated by Table 3, students' final draft scores over 7 terms (Spring 2012 to Spring 2014) of first-year composition, both 1101 and 1102, were analyzed first. The final projects were used because those scores reflected the best efforts: these drafts were written after students received feedback from their instructor on an early and intermediate draft, had a face-to-face meeting with the instructor for two of the three projects, and received feedback from their peers, typically at least three peers but often five peers.

Table 3
Sample Sizes of Final Drafts

Class	Term	N
ENC 1101	Spring 2012	351
ENC 1101	Summer 2012	648
ENC 1101	Fall 2012	3088
ENC 1101	Spring 2013	525
ENC 1101	Summer 2013	96
ENC 1101	Fall 2013	4114
ENC 1101	Spring 2014	524
ENC 1102	Spring 2012	3347
ENC 1102	Summer 2012	306
ENC 1102	Fall 2012	2943
ENC 1102	Spring 2013	3655
ENC 1102	Summer 2013	226
ENC 1102	Fall 2013	2635
ENC 1102	Spring 2014	4443

The average rubric scores (Ravg is obtained by adding together the eight subscores and dividing by eight) for each scored paper creates a mound-shaped distribution that ends abruptly on the right. This clipping off of the distribution represents information lost when instructors have reached the maximum score assignable. We notice and note this due to its impact on the analysis of our data.

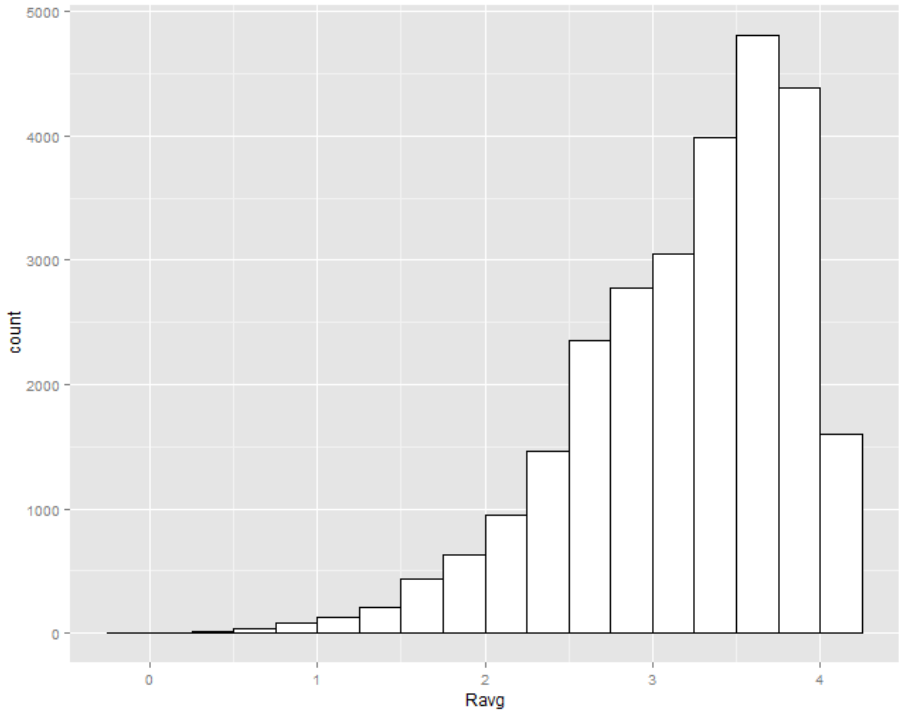

Figure 1. Distribution of all Ravg ratings for final drafts, bin size = .25

As a result of the clipping seen in figure 1, information about the high-scoring students is lost: once they reach the upper limit of four, further increase cannot be measured. Ideally, the right side of the distribution would more resemble the left, but this would require a scale of approximately 0 to 6 to accommodate the symmetrical shape. Such clipping has negative effects on the regression analysis.

A relatively small number of students (310) took a course more than once. These duplicates were removed in order to study only ordinary course completion. Additionally, any scored assignment for which the Ravg was zero was eliminated because these generally indicate assignments that were not completed. Data merging and statistical analysis was done with R version 3.3.1 (R Core Team 2014). Where error bars appear in figures, they represent two standard errors of the estimate of the mean.

4. Results

Considering the use of rubric-assisted instructor evaluations as a means of programmatic assessment requires testing both the construction of the rubric and the role of the evaluator in the scoring. To evaluate the impact of both the rubric and the instructor on the score, our research investigates subscores and their relation to average rubric ratings, examines the correlation between subscores, and explores instructor scoring over two terms.

4.1 Construct validity of the rubric

To test the construct validity of the rubric, the subscores and their relationship to average rubric rating were examined. Using a principle components analysis, which attempts to measure

how independent the pooled 1101 and 1102 subscores are from one another, the relationships between subscores are analyzed by decomposing them into perpendicular dimensions with the expectation of seeing at least two or three of these dimensions well-represented, as measured by the variance captured.

Table 4
Principle components analysis

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
%Var	.58	.11	.09	.06	.05	.04	.04	.03
R1	0.340	0.395	0.066	0.775	0.180	0.153	-0.255	0.054
R2	0.380	0.224	0.262	0.042	-0.110	-0.610	0.589	-0.074
R3	0.361	0.261	-0.025	-0.209	-0.802	0.308	-0.136	0.009
R4	0.385	0.031	0.241	-0.344	0.389	0.396	0.211	0.568
R5	0.388	0.059	0.217	-0.382	0.323	-0.083	-0.441	-0.590
R6	0.334	-0.611	-0.079	0.255	-0.034	0.384	0.377	-0.390
R7	0.348	-0.561	-0.030	0.084	-0.136	-0.433	-0.426	0.411
R8	0.280	0.181	-0.902	-0.135	0.192	-0.108	0.087	0.012

In Table 4, the first component (the dimension with the most explanatory power) captures 58% of the variance of the subscores while the others are considerably smaller, which indicates that the scores do not distinguish clearly the putative categories used in their descriptions. The scores are conflated either by the nature of what is being measured or by the raters themselves. Either way, the finding challenges the construct validity of the eight subscores and calls into question the meaningfulness of distinguishing between them.

As a corollary to this finding, the first factor (the first column) is populated with nearly identical numbers, which means that the most important aspect of the subscores is that taken together they behave much like an average (with a slight underweighting for R8). As noted, this dimension accounts for 58% of the variance in scores. Although component two begins to distinguish between traits Style and the combination of Focus and Evidence, the support for construct validity of the traits is weak and overwhelmed by the average rating for the paper over the eight subscores. Fortunately, averaging the subscores is the way the grade for a paper is computed, so the primary component corresponds directly to Ravg and the grade.

While the finding represented here challenges the validity of the rubric construction, it does simplify the rest of the analysis by justifying the examination of the properties of the average of the eight subscores as a proxy for instructor's measurement of overall student writing and development; it also greatly simplifies the presentation and streamlines the analysis to one index (the average of the subscores) rather than eight.

4.2 Instructor scoring and predictive validity

The construction of the two-term course sequence (1101 and 1102) allows the average scores of students in 1102 to be modeled using what is known about 1101; this is accomplished

by creating two independent statistics from 1101: one for each student and one for each instructor. For this purpose, only data from instructors and students who were in both 1101 and 1102 is used, and students who had the same instructor for both courses are eliminated. For each of these, a "random effect" is calculated as the average rating assigned to (for students) or by (for instructors) minus the average rating assigned in 1101 by that student's instructor. For example, if the student effect is .5, it means that his or her average received rating in 1101 was .5 more than the average for all students who had the same instructor. Similarly, an instructor effect of .5 means he or she *assigned* scores that were on average .5 higher than average over all instructors. In this way, any student's score is the [average over all instructors] + [instructor effect] + [student effect]. A statistical model of this was built as a multiple regression on these two random effects to predict 1102 ratings of each student. The fit of the model is a measure of the reliability of the ratings.

Table 5

Summary of Regression Analyses for Variables Predicting Average 1102 scores from 1101, N = 1486

Variable	*B*	*SE B*	*β*
Intercept	2.91	.01	.00
Instructor Effect	1.50	.06	.57
Student Effect	0.43	.02	.43
R^2		.29	
F		309	

$p < .001$ for each estimate

The p-values have to be taken with a grain of salt because the residuals are not normally distributed; this does not invalidate the model coefficients, but it overstates the significance. The model captures 29% of the variance. The instructor effect coefficient is greater than one, which means that instructor rating habits tend to persist from one term to the next. By contrast only 43% of the student effect persists with the new instructor. The correlation of instructor average assigned scores between 1101 and 1102, using the same data as the model, is .74.

Another way to investigate the relationship between 1101 and 1102 is to examine student rankings between early and late projects. The second 1101 project was used instead of the first because of the low score averages on the first project, which suggests that students and instructors are still adjusting and that score may not be representative of their capabilities. There were 2,209 students who had both final draft scores to compare.

Table 6

The transition frequencies between the second 1101 project and the last 1102 project

	1102 lowest 25%			1102 highest 25%
1101 lowest 25%	229 (41%)	166 (30%)	94 (17%)	64 (12%)

	150 (27%)	154 (28%)	136 (25%)	112 (20%)
	116 (21%)	124 (23%)	172 (31%)	140 (25%)
1101 highest 25%	58 (10%)	108 (20%)	150 (27%)	236 (43%)

Table 6 shows the numbers in each cell, as well as the (rounded) transition frequencies, based on rows that sum to 100% (552 or 553 cases, where ties forced some cases arbitrarily into a higher or lower quartile). For example, among students in the lowest quartile of 1101 project 2, only 12% finished in the top quartile of 1102 project 3.

If there were no persistent student traits associated with earning good ratings on writing, each cell would be expected to be 25% in order to show complete random sorting over time. Table 6 proves that this is not the case. The values greater than 25% indicate a tendency to retain the initial ranking, and lower than 25% means the transition is rarer than chance; this was formally tested with a Chi-Squared test of independence on the cell counts, $X(2, N = 16) = 263.8, p < .001$, which rejects the hypotheses that the two quartile rankings are independent.

4.3 Scoring and sequencing

The intersection of the rubric and the instructor results in the score. The course sequence allows scores to be examined over two semesters for changes that impact the potential of the scores to be used for assessment beyond the course and the classroom. Recall that the Ravg average score for each set of scores is obtained by simply adding together the eight subscores and dividing by eight. Because 1,887 students who took 1101 and 1102 sequentially, changes in scores over the two semester sequence can be examined. Students receive different instruction and different writing prompts in each class, but both have three projects each and use the same rubric for scoring.

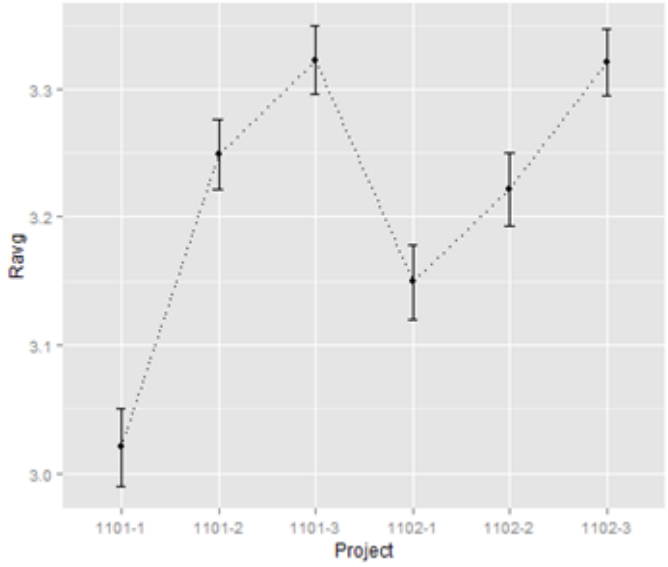

Figure 2. Ravg averages over time for common completers of 1101 and 1102

The two-semester history of average scores on final student papers shown in Figure 2 demonstrates rising scores within each of the two courses with a dip between them. The dip will be referred to as a *reset effect.* On average, students show a score reset between 1101 and 1102

with average scores at the end of 1101 about .16 rubric points higher than at the beginning of 1102. As a result, the average scores of the third project in both courses are about the same.

Generally, 1101 students are placed into that class because they are judged (via placement tests or prior academic records) to need more development than *native* 1102 students. Therefore, if 1101 has no permanent effect on them, their 1102 scores would be expected to lower than students directly placed into that class. On the contrary, the opposite is true, as a t-test of difference of means shows. The students who completed 1101 and then took 1102 show about a .1 bonus over native 1102 students (p-value < .001), or more than a one-project lead in development.

5. Discussion

The results suggest significant implications for our program and for the value of instructor scores for programmatic assessment. The findings demonstrated in 4.1 challenge the construct validity of the rubric and suggest that the eight subscores are being conflated. The findings illustrated in 4.2 demonstrate that both instructor grading styles and student traits are somewhat persistent across time and that scores are predictive. Finally, findings discussed in 4.3 reveal the reset effect and support the two-term sequence. Interpretations collapse categories but can be generalized into discussions regarding the rubric, the instructors, the scores, and the limitations and extrapolated as they impact and interact with ongoing scholarly conversations.

5.1 Rubric

The findings demonstrated in 4.1 challenge the construct validity of the rubric and suggest that the eight subscores are being conflated either by the instructor application or as a result of the category construction. Taken together, the subscores behave much like an average, and since averaging the subscores is how the grade for a paper is computed, that means whether by construction or application, the rubric is functioning primarily as a holistic grade. Analysis does not refute rubric use in general but does suggests that the limiting factors resulting from the construction of the rubric in consideration warrant attention, specifically with regard to the scoring scale and subscore structure.

An obvious issue with the rubric is that the range of the scoring scale is not large enough to accommodate growth over a semester, much less over a two-semester sequence. Figure 1 illustrates that the 0 to 4 scale is used predominantly at the upper end, creating a skewed and clipped distribution. Because the principle component analysis depends on the pairwise variation in subscores, the limited range of the scale undoubtedly degrades the ability to identify the qualitative distinctions desirable in the dimensions shown in Table 4. The reset effect observed between 1101 and 1102 (Figure 2) reinforces this interpretation.

Also, according to the principle components analysis (Table 4), since 58% of subscore variance constitutes the average score, requiring instructors to click eight rubric criterion seems counterproductive. The instructors rarely award zeros or ones, which results in a cluster of grades at the high end. Anecdotally, some instructors admit *reverse rating* by assessing a paper holistically, assigning the appropriate grade, and then choosing suitable subscores to make the average match, which allows a student to locate areas for potential improvement and provides the instructor an explanation or justification of the score. While it is impossible to know the extent to which reverse rating is employed without additional research, the scores suggest that it

is reasonable to admit that instructors are not capitalizing on the stratification of the eight rubric subscores. An initial conclusion from the forgoing discussion is that the scale in use is too limited to reveal the qualitative distinctions the design of the rubric is intended to measure.

Furthermore, the subscore structure intended to distinguish between Basics and Critical Thinking appears to be ineffectual. The correlation between R1-R2, R4-R5, and R6-R7 suggests that the Basics and Critical Thinking pairs are being treated as a single subscore, which makes it seem reasonable to include both levels of a trait in a single subscore in conjunction with extending the range beyond the 0 to 4 scale—that is, to have fewer subscores but allow for more levels of discrimination within each. A similar statement can be made regarding the five traits the rubric is designed to measure (Focus, Evidence, Organization, Style, and Format). Table 4 shows that there is little support for construct validity of these traits in the data.

In retrospect, since the distinction between Basics and Critical Thinking was included on the suggestion and assumption that it would support institutional and external assessment, finding no distinction, while a result in and of itself, necessitates a lack of need for the distinction. Additionally, the finding could imply an actual lack of distinction, an argument that would reinforce Washington State University's finding that "no relationship existed between our writing assessment scores and our critical thinking scores" (Condon & Kelly-Riley, 2004, p. 63). Either interpretation argues against the need for a distinction.

The findings related to the construct validity of our rubric reinforce Rezaei and Lovorn's (2010) claim that raters tend to ignore rubrics, but by reinforcing arguments of subjectivity within rubric usage, the findings refute claims that the standardized responses of rubrics undermine what is should be a subjective activity (Elbow, 2006; Wilson, 2007). Implications connect to larger conversations regarding rubric usage by suggesting that caution is necessary and reflection is required, as is the case in all teaching and grading practices, but that analytics allow problems to be identified and corrections to be applied and that even the big, generic rubrics used across terms and genres that many in Writing Studies fear most allow for flexibility in instructor application. Tools may impact instructor approaches, but they do not define them.

5.2 Instructors

The findings illustrated in 4.2 demonstrate that instructor grading styles are somewhat persistent across time. Additionally, analyzing instructor scores establishes that student traits exist across two semesters and students maintain quartile rankings. While this finding also establishes that scores are predictive, which is useful for intervention purposes, it confirms the impact of instructor style on student scores.

If scores have some validity to compare competencies within 1101 and 1102 separately, then the relative ranking of students in each of the classes will be retained to the extent that student characteristics persist and the grading styles of instructors persist. The model in Table 5 finds that 29% of the variance in average 1102 scores was captured in the 1101 data. By inspecting the model coefficients, it becomes clear that the instructor effect is preserved at about 100%, while the weight of the student effect is only 43%. Ideally (or perhaps idealistically), the instructor effect would not be significant, and the student effect would be 100%. The influence of the instructor effect over the grade is an undesirable source of variance in rubric scores; it reduces inter-rater agreement and indicates that grades depend partly on the grading habits of their instructor.

Despite the impact of instructor scoring style, student performance is still very important, as demonstrated in Table 6: past performance predicts future performance in a student's ranking relative to others. Quartile persistence raises interesting questions and presents the possibility that narratives of failure might result from quartile maintenance. Students who persist at the bottom quartile might receive reinforcing discourses of discouragement, and students restricted from score increase due to the four-point scale censoring may also be impacted by what they consider an inability to improve, which suggests that instructor feedback by quartile should be analyzed carefully. Even if the individual dimensions of the rubric cannot be entirely trusted, evidence for score reliability argues that the rubric-assisted scores are measuring *something*.

Recognizing the impact of instructor grading style on scoring in conjunction with the finding that rubrics are applied in context allows program administrators to capitalize on this by creating and expanding contextual analytics that allow instructors to recognize where their grading style is located in relation to their specific class and other instructors. Instructors can be provided with real-time analytics that place their scoring in the context of all instructors' scores on the same assignments, which could alert instructors if they are outliers. Outside factors could be considered, as well. In all, significant opportunities for improvement and innovation can be established by contemplating the findings generated by analyzing instructors' scoring.

General suspicion of the evaluation of writing is grounded in the inherent subjectivity of the evaluation process, and our findings support the presence of subjectivity across evaluators but argue against complete subjectivity of individual instructor evaluations. The presence of stable instructor scoring styles rebuts assumptions that instructors display subjectivity from student to student in their evaluations but reinforces claims of subjectivity from teacher to teacher. Such findings bolster claims that instructor feedback remains subjective despite the application of a rubric (Gentile, 2000) but discredit the solution of leaving the evaluation of writing to subject-matter experts because that would do nothing to alleviate the variation from teacher to teacher across disciplines and subject matters. Further, such differences can be assumed beyond the evaluation of writing given that no degree of systemization can mitigate the differences inherent in the instruction and evaluation of different instructors. Nevertheless, the presence of subjectivity in the face of attempts at systemization again refutes generalized fear of rubrics.

5.3 Scores

Finally, examining scores over two semesters, as tested in 4.3, supports the construction of the two-term sequence and establishes the existence of a reset effect. Despite being categorized as requiring more development, students who took 1101 scored higher in 1102 than native students in 1102, which suggests a transfer of learning. The reset effect, however, suggests an inherent conflict of interests in using these scores in programmatic assessment.

The reset effect (Figure 2) signals the reset of scoring based on the new classroom or what we will call the *rater culture*, which suggests that each class has an individual context. Essentially, each class resets the accomplishment scale back to the beginning to allow for evidence of improvement over the semester. The reset effect indicates two related phenomena. First, within each of the 1101 and 1102 courses, average ratings increase over time across the three assignments. If this were not the case, it would indicate that either students were not learning or that instructors were not measuring their competency. As it is, the growth in scores (as supported by the predictive model) is a tentative affirmation that within a semester, the rubric

has some power to describe the quality of student work. The second phenomenon is the reset from higher scores at the end of the 1101 course that results in average scores for the first project in 1102 that are lower than the second project in 1101. If the rubric scores tracked growing competency, this reversal of progress would not be expected. Instead, it appears that the rubric scores are recalibrating for the second course, which implies that the ratings are not directly comparable. The inability to show incremental progress over time is an argument against using the rubric scores from either course as an absolute measure of competency.

Multiple potential causes of this reset are supportable. Perhaps, as suggested by Goldman and Hewitt (1975), faculty set grading expectations according to the perceived capabilities of students; it is possible, then, that our instructors tend to use the rubric contextually by setting a class standard that is suitable to demonstrating growth over a term rather than an absolute (reified) scale. If the scoring for the first project in 1102 began at an average of 3.2 (where it ends in 1101), there would be almost no way to show improvement across the semester. Reinforcing the implications discussed in 5.2, this supports the likelihood that the analytic rubric scale provided is being applied pragmatically—if in an *ad hoc* manner—by individual instructors.

Another possibility is that instructors start out *tougher* in relation to their idea of scoring, perhaps to establish authority or guard against inflation, and then ease up over the course of the term, resulting in an arbitrary substantiation of development across the course. An alternative explanation is that students must learn to write to a new audience at the beginning of any course, and their performance score resets as they learn this new audience. Or the reset may simply illustrate the complexity of writing development, as illuminated by researchers in Writing Studies who have found the ability to write well is not mastered in one context and then simply carried over to another context (Bazerman, 1988; Beaufort, 2007; Carroll, 2002; Nowacek, 2011). Regardless of cause(s), it is worth noting that the reset effect is less evident for those students who come to 1102 directly from 1101. Despite presumably starting with less writing competency, the fact that they outscore students who place directly into 1102 on the first project reinforces the implication that knowledge is transferring.

Given the context of how the rubric scores are used, the reset is inevitable; the average score is calculated to become the grade for the project, and because of the way that course grades are generally apportioned, the rubric scores will naturally align with grading practices. By contrast, if the rubric scores reflected increasing student development over two semesters, it would by definition force very low grades in the first semester. As a result, forcing the direct numerical correspondence between rubric scores and assigned grades forces the reset, which means any absolute measure of progress from 1101 to 1102 is lost.

The reset effect suggests the establishment of a rater culture that allows the rubric to function in context but also demonstrate a lack of reliability that discourages using these scores as an absolute measure. While evidence of contextual use should allay fears that rubrics eliminate instructor agency or autonomy, the impact of instructor scoring style and the required tailoring of rubric application to course-specific needs and outcomes supports and strengthens the argument that instructors' classroom scores are not a reliable means of assessing programs. Tying programmatic assessment to course grades can be useful because it transparently communicates the values of the rater in a standardized framework, but the reset effect demonstrates that there is an inherent conflict of interests in using ratings for these dual purposes, namely that it forces the scope of the rubric ratings to be tied to the scope of the course in which they are used. Decoupling assessment from grading would allow the potential for a

rubric scale that spanned the two-course sequence, which could improve the value of the assessment data as a longitudinal measure of student competencies. Alternately, using the analytics afforded from digital capture of scoring and commenting allows for other avenues of immediate intervention such as the inclusion of multiple raters or options such as scoring across sections, among other possibilities. Further consideration could be given to the use of developmental rubrics instead of analytic rubrics, as well.

5.4 Limitations

Some limitations of the research and its findings are evident. First, the limited scale on the analytic rubric combined with rater habits produces a distribution of scores that are not ideal for analysis or pedagogy. The limited variance in scores reduces the ability to detect factors that might validate the individual rubric traits or the categories into which they are grouped. Additionally, ratings at the upper (clipped) end of the distribution do not allow levels of excellent work to be distinguish since papers cannot receive a rating greater than four. Also, while the regression model and the quantile transition table support the meaningfulness of the data with respect to measuring student competency in that they are consistent with patterns that would be expect for development of student competencies, and they have some predictive power, there is no trusted criterion to test rubric scores, such as high quality ratings that are independently rated by multiple reviewers, which leaves some important validity questions unanswerable at the moment.

6. Conclusion

The application of digital tools and the resources they create have expanded the possibilities surrounding evaluation and assessment. As Writing Program Administrators, we have supported and participated in the use of rubric-assisted, classroom-based instructor scores as means of programmatic assessment and can confirm that the rewards of this practice are as clear and considerable as they appear. As researchers, however, our findings suggest that the concerns with such practices cannot be ignored. Utilizing instructor-generated classroom scores to assess overall changes in student scores in search of improvement as an indication of learning appears viable. Equally, assessing instructors, course material and structure, and the course rubric result in usable data. But while instructor scores provide valuable measurements for internal programmatic assessment, decontextualization degrades their analytic accuracy.

Program administrators are likely to have an understanding of the rhetorical situation in which the scores were constructed and be capable of contextual interpretation, but in light of the influence of the classroom context on score construction, every subsequent step away from the source of the score weakens their evaluative validity. Arguably the farther program administrators are removed from the classroom and the curriculum, the less effective they will be at analyzing the scores, as well. For internal programmatic assessment, including individuals who create and convey the curriculum in the assessment process is necessary to provide context for an interpretation of the scores and of the assessment. Extending the scope and extrapolating the interpretation of such evaluations to include large-scale contexts beyond the program or involving university administrators and outside accrediting bodies or external stakeholders in the evaluation of highly contextualized outcomes devalues the usefulness and usability of the scores.

That said, programmatic assessment that is of the program, by the program, and for the program possesses the inherent limitations and potential conflicts present in all forms of self-assessment.

There is no question that the use of instructor ratings of student writing, especially digitally-captured scoring, is an invaluable resource, and ideally, these findings will incentivize considerations regarding how to deconstruct and reconstruct scores for application beyond courses and course grades. Data sets can be culled or cleaned; perhaps instructors whose scores are defined as outliers based on standard deviation could be clipped. And data can be narrativized. Possibly viable external-internal assessment, ideally external to the classroom and perhaps program but internal to the college or campus, could encompass all departments but would remain constricted to maintain a generally sustained student environment and to allow access to instructors who could share contextualizing stories. Similarly, digital tools could be used across the curriculum and in the discipline in designated writing intensive courses to aggravate e-portfolios that trace student development and transfer across a college career. If writing assessment has progressed from ideals and aims of objectivity and standardization and moved through holistic assessment to portfolio evaluation, all the while balanced on the defining concepts of validity and reliability (Yancey, 1999, p. 484), a logical next step would be to add an *e* to portfolios and expand the available data in order to test the validity and reliability of e-portfolios.

Ultimately, our findings have mixed results for the use of rubric-assisted instructor scores for programmatic assessment. Pragmatically, the use of instructor evaluations for programmatic assessment depends on the strength of the instructor evaluations, which result from the alchemical merger of instructor, class, curriculum, and tools—in other words, the strength of the instructor rating is related directly to the strength of the program. Programmatic assessment is reciprocal, then, in that it improves programs and improves instructors and instruction, which improves the viability of accessing instructor ratings for programmatic assessment. In this way, the valuable program-related information our study provided will improve the usefulness of future scores for future programmatic assessment.

In the larger context, the results of this analysis suggest that major arguments supporting the distrust of rubrics are unsubstantiated, and the distrust of instructor evaluations of writing seems overstated but complicated. The distrust of scores, specifically in their direct application for programmatic assessment, appears justified, at least for cases of extended or external assessments. The application of digital tools and the production of writing analytics have the potential to create transparency and facilitate discussions that promote trust; currently, however, the warning of the CCCC Committee on Assessment against indirect assessment and the cautioning of the National Council of Teachers of English and the Council of Writing Program Administrators against decontextualization appear warranted. While the potential identified by considering the use of digitally-captured, rubric-assisted, classroom-based instructors' scores in relation to the guidelines offered by these disciplinary bodies remains, the findings of this study suggest that limiting the extension of classroom evaluations is justified.

What remains unsettled is the struggle to reconcile the subjective through standardization. Even given the degree of systemization present in the data set under examination, standardization was unsuccessful and subjectivity unavoidable. In ways, Writing Studies accepts this absence of objectivity by fighting for individualism and subjective subjectness; in other ways, however, the discipline seeks the simplicity of right answers. The tension between the quantifiable and the qualitative undergirds questions of assessment that drive instructors insane from attempts to define quality (Pirsig, 2006). As technology continues to expand and advance and theories form

to understand its uses and consequences, current resistance to overgeneralization and extended extrapolation will lay a foundation to guard future students and instructors of writing from yet-unknown dangers and dilemmas. But assessment must advance. And just as concepts around the construction of writing evolved from writing as product to linear process to complex and induvial processes, perhaps the nuances that separate what can be learned and assessed from what cannot be taught or calculated will simply become part of a percentage we reserve for finesse.

References

Anson, C., Dannels, D., Flash, P., & Housley Gaffney, A. (2012). Big rubrics and weird genres: The futility of using generic assessment tools across diverse instructional contexts. *Journal of Writing Assessment, 5*(1), n. pag.

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

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

Conference on College Composition and Communication. (2014). "Writing Assessment: A Position Statement." CCCC Committee on Assessment. Retrieved from http://www.ncte.org/cccc/resources/positions/writingassessment

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

Caldwell, O. W., Courtis, S. A., Courtis, S. A., & Courtis, S. A. (1971). Then & now in education, 1845: 1923; a message of encouragement from the past to the present. *Ebscohost*. New York, Arno Press, 1971 [c1923].

Carter, M. (2003). What is the difference between assessing a program and assessing a student? Retrieved from http://www.ncsu.edu/provost/academic_programs/uapr/FAQ/UAPRFAQ whatdifassessstudentvsprograms.html, 2003.

Condon, W., & Kelly-Riley, D. (2004). Assessing and Teaching What We Value: The Relationship between College-Level Writing and Critical Thinking Abilities. *Assessing Writing*, *9*(1), p56-75.

Council of Writing Program Administrators. (2008). NCTE-WPA White Paper on Writing Assessment in Colleges and Universities. *CWPA*. Retrieved from http://wpacouncil.org/whitepaper

Cressey, P. F.. (1929). The Influence of the Literary Examination System on the Development of Chinese Civilization. *American Journal of Sociology*, 35(2), 250–262. Retrieved from http://www.jstor.org/stable/2766126

Cureton, L. W. (1971). The History of Grading Practices. *National Council on Measurement in Education, 2*(4), 1-8. Retrieved from http://tenure.umatter2.us/wp-content/uploads/2011/09/1971-Curreton-History-of-Grading.pdf

Elliot, N. (2005). *On a Scale: A Social History of Writing Assessment in America.* New York: Peter Lang. Print.

Elbow, P. (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(2), 81–99.

Gentile, J. R. (2000). An Exercise in Unreliability. *Teaching of Psychology,* 27(3), 210.

Gerretson, H., & Golson, E. (2005). Synopsis of the Use of Course-Embedded Assessment in a Medium Sized Public University's General Education Program. *JGE: The journal of general education*. Vol. 54, No. 2, Retrieved from https://www.bmcc.cuny.edu/iresearch/upload/CourseEmbeddedAssessment.pdf

Goldman, R. D., & Hewitt, B. N. (1975). Adaptation-level as an explanation for differential standards in college grading. *Journal of Educational Measurement,* 12(3), 149-161.

Huot, B., O'Neill, P., & Moore C. (2010). A usable past for writing assessment. *College English*, 72(5), 495-517.

Inoue, A. B. (2005). Community-based assessment pedagogy. *Assessing Writing,* 9(3) 208-238.

Inoue, A. B. (2014). Theorizing failure in US writing assessments. *Research in the Teaching of English,* 48(3), 330-352.

Langbehn, K., McIntyre, M., & Moxley, J. M. (2013). Re-mediating writing program assessment. In H. Mckee, & D. N. DeVoss (Eds.), *Digital writing assessment & evaluation*. Utah State University: Computers and Composition Digital Press.

Mason, G., & Dragovich, J. (2010). Program Assessment and Evaluation Using Student Grades Obtained on Outcome-Related Course Learning Objectives. *Journal of Professional Issues in Engineering Education and Practice*. October. DOI: 10.1061/ASCEEI.1943-5541.0000029

Moxley, J. M. (2013). Big data, learning analytics, and social assessment methods. *Journal of Writing Assessment, 6*(1), Retrieved from http://www.journalofwritingassessment.org/article.php?article=68

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

Plutarch. *Nicias*. Bernadotte Perrin (Ed.). Perseus Digital Library. Retrieved from http://data.perseus.org/citations/urn:cts:greekLit:tlg0007.tlg038.perseus-eng1:29.2

Pirsig, R. M. (2006). *Zen and the art of motorcycle maintenance: An inquiry into values*. New York: HarperTorch.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Rezaei, A. R. & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing* 15 (1.) 18–39.

Rogers, G. (2003). "Do grades make the grade for program assessment." Retrieved from http://www.abet.org/wp-content/uploads/2015/04/do-grades-make-the-grade.pdf

Sax, G. (1980). *Principles of educational and psychological measurement and evaluation* (2nd ed.). Belmont, CA: Wadsworth.

Sommers, N. (1982). Responding to student writing. *College Composition and Communication, 33*(2), 148-156.

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

Starch, D., & Elliott, E. (1913a). Reliability of grading work in history. *School Review*, 21, 678-681.

Starch, D., & Elliott, E. (1913b). Reliability of grading work in mathematics. *School Review*, 21, 254-295.

Swearingen C. J. (1991) *Rhetoric and Irony, Western Literacy and Western Lies,* Oxford University Press.

Thayer, V. T. (1928). *The passing of recitation.* New York, Heath Publishing.

Turley, E. D., & Gallagher, C. W. (2008). On the uses of rubrics: Reframing the great rubric debate. *English Journal*, 97(4), 87-92.

Vieregge, Q., Stedman, K., Taylor Mitchell, J., & Moxley, J. (2012). *Agency in the Age of Peer Production*. Urbana, IL: National Council of Teachers of English.

Willingham, W. W., Pollack, J.M., & Lewis, C., Educational Testing Service. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*. Spring 2002, Vol. 39, No. I , pp. 1-37.

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, K. (1999). Looking back as we look forward: historicizing writing assessment. *College Composition and Communication*, Vol. 50, No. 3, pp. 483-503.