Helping Faculty Self-Regulate Emotional Responses in Writing Assessment: Use of an "Overall Response" Rubric Category

by Michelle E. Neely, University of Colorado Colorado Springs

Faculty evaluation of student learning artifacts is a critical activity as accrediting bodies call for campuses to promote "cultures of assessment." Also important are opportunities for faculty engagement and development that assessment projects provide. However, such projects come with significant challenges to facilitators and faculty scorers themselves. Faculty bring their own expertise and beliefs about student learning and writing to the assessment context, all of which can have emotional valence. Assessment sessions may emphasize faculty scoring to components of a rubric, perhaps eliminating a holistic score from the conversations because holistic scores are not viewed as actionable data points and thus are often not parts of the rubric (McConnell & Rhodes, 2017). As actionable data and reliability among scorers are emphasized in assessment, and holistic scores fall away, are we losing an important scoring tool by removing a place for assessment scorers to log their overall responses to the work that they are evaluating? This exploratory study reports scorers' use of an "overall response" category, added to the rubric in two assessment projects. Results indicate that faculty found the new category helpful, not just in managing their emotional responses, but also in leveraging their emotions to complete the scoring task. Correlation and regression analyses also suggest that scorers maintained orthogonal scoring across rubric categories.

The Association of American Colleges and Universities' (McConnell & Rhodes, 2017) report, *On Solid Ground*, emphasized faculty engagement in assessment processes, to include use of authentic student artifacts and cultivation of teacher expertise as evaluators of student work. Findings suggested rubrics can be used reliably to assess written communication and critical thinking and are useful as points of conversation around faculty values and shared learning outcomes across disciplines. With regard to writing assessment specifically, recent scholarship by Condon, Iverson, Manduca, Rutz, and Willett (2016) explored university writing assessment as a "site of faculty learning" (p. 53), reporting that, across disciplines, faculty who participated in writing portfolio assessment reflected upon and shifted their own teaching practices around assigning writing based on their assessment experiences. These findings have implications for the role of assessment activities as facilitating teacher development and change (Guskey, 2002).

The task of reading student artifacts against a rubric, with an eye towards doing so efficiently and reliably with other scorers, is challenging. Assessment scoring is evaluative, but it is not the same kind of evaluation that one does for a piece of student work in a class that one teaches; thus, it is a different skill for teachers to develop. Yet, the skill is valuable as accrediting bodies call for campuses to promote "cultures of assessment" (Ikenberry & Kuh, 2015) and seek evidence of such initiatives. Perhaps more importantly though, these types of assessment projects provide critical faculty engagement opportunities for individuals to articulate and negotiate their values around student learning, revisit programmatic goals, and learn about the practices of their colleagues (Broad, 2003; Condon et al. 2016; Huot, 1997; Walvoord, 1997).

Operationalizing authentic, faculty-led assessment practices can also be complex for facilitators. Whether using rubrics or taking a more comprehensive and organic approach of Dynamic Criteria Mapping (Broad, 2009), assessment tasks involve groups of faculty reading student papers and discussing their responses. These sessions take place for hours at a time generally across several days (Dryer & Peckham, 2014; Huot, 1997). Often coordinated by writing program administrators or assessment leaders, facilitators work to help faculty scorers set aside their own grading practices and notions of student writing so they can contribute to the task at hand: evaluate evidence of student learning to improve teaching and educational practices (Kuh & Hutchings, 2015).

The potential for faculty—and programmatic—change is perhaps what makes this assessment work so valuable and exciting (Condon et al., 2016; Walvoord, 1997). Discussing student learning artifacts, usually student papers, involves a type of unique "seeing" of student learning, different from the type of reading that faculty do to determine student grades in their courses, where they have knowledge of individual students' trajectories, effort, and class contexts (Eckes, 2008). Assessment conversations touch at the very core of faculty members' beliefs about teaching and learning (Broad, 2003). At their best, they involve faculty values and epistemologies, and as a result the participants may find themselves working to manage emotional responses within the task of student paper evaluation and calibration discussions with colleagues.

Assessment sessions generally emphasize faculty scoring to *components* of a rubric, perhaps even eliminating a holistic score from the conversations (Hamp-Lyons, 2016), because holistic scores are not part of the rubric (McConnell & Rhodes, 2017) and do not produce actionable data. As actionable data and reliability among scorers are emphasized in assessment practices, and holistic scores fall away, are we losing an important scoring tool by removing this place for assessment scorers to log their overall responses to the work that they are evaluating?

In the exploratory study reported here, I work to address this issue of the internal usefulness of the "holistic" category and its potential function as a space for faculty scorer emotional regulation. Prior work (Caswell, 2014; Edgington, 2005) has demonstrated faculty members' emotional responses during their grading of their own students' writing, where emotions were understood as

providing a type of feedback (Turner, 2010) that course faculty could use to make direct changes to assignments, classroom instruction, and other student support based on their experience of evaluating their students' papers. However, the context is different for faculty scorers participating in assessment contexts, where faculty read artifacts written by students in different courses, and sometimes even from previous semesters, in the case of larger-scale programmatic assessments (e.g., Haswell, 2000). Thus, assessment contexts provide a different challenge for us to understand and explore faculty members' emotional responses.

To better understand how faculty scorers may experience emotions in assessment scoring contexts, I first provide an overview of work about teacher beliefs and their influence on practice, strategies used to provide scorer support within group contexts, and the influence of teachers' emotions on evaluation tasks. I then provide some definitions of emotions and emotional coherence as they may relate to independent, orthogonal scoring, and scholarship on the rhetoric of emotion. Finally, I report on the use of an "overall response" rubric category added to the rubric provided to scorers during two recent assessment projects as a means to ease some of these challenges that faculty writing assessment scorers may face.

Faculty Beliefs and Highly Personalized Pedagogies

Meaningful assessment of student learning is that which is local and contextualized, and there is exigency to answer the call for localized assessment because, as Broad (2009) reminded us, "Commercial testing corporations are eagerly inviting us to outsource" (p. 2) our writing assessment to them, providing an array of testing formats for our campuses to purchase, which would allow their products to tell the story of student learning at our institutions. Operationalizing local assessment, however, is messy work. While faculty expertise is the very lens through which we want assessment artifacts evaluated, this expertise comes with all of the quirks, personalities, and biases of individual teachers as they learn to read student writing for the purposes of assessment. In short, faculty experience is bound up in idiosyncrasies and preconceptions that assessment facilitators must help individuals and groups navigate for assessment projects to be successful.

Donna Kagan (1992) used the term "highly personalized pedagogy" to describe the personalized and individualized nature of beliefs informing teachers' practice. About this, she explained, "As a teacher's experience in the classroom grows, his or her professional knowledge grows richer and more coherent, forming a highly personalized pedagogy—a belief system that constrains the teacher's perception, judgment, and behavior" (p. 74). Thus, faculty with deeper sets of experience may also have sets of beliefs that are more restrictive of their judgments. Faculty beliefs are generally assumed to inform their teaching practices (Fives & Buehl, 2012; Richardson, 1990), have strong affective components often informed by teachers' own experiences as students themselves (Neely, 2017; Strømsø & Bråten, 2011), and are resistant to change (Pajares, 1992). In other words, faculty members each hold personalized pedagogies, or beliefs, that run deep, have emotional valence, and are relatively intractable, and they bring these to the context of assessment work, just as they bring them to other professional tasks.

So, while we want faculty to engage in assessment work, to add meaning to results and as professional development opportunities, faculty each come with their own set of unique, highly personalized pedagogies that may inform their evaluations of the assessment artifacts. These biases may serve as internal constraints within the faculty evaluator, as s/he works to evaluate student work. They may also constrain within group practices as the evaluation group works to come to consensus during rubric discussions, during which time the groups of faculty assessors calibrate their evaluations to rubrics or other assessment guides, negotiating meaning and shared values with fellow raters.

Group Process Support in Assessment Work

Work by Colombini and McBride (2012) explored raters' conversations in scoring sessions and the specific ways that dissent among scorers provided opportunities for group growth relative to group dynamic theory. Emphasizing faculty engagement and critical reflection in assessment practices over psychometric reliability (Charney, 1984; Huot, 2002), Adler-Kassner & O'Neill (2010) explained the potential of both "conflict and consensus as dual goals" (p. 192) in facilitating writing assessment. Similar work by Jølle (2014) also examined dialogue among faculty raters, specifically pairs of novice scorers, as they completed a scoring project. They found that rater conversations about the scoring activity and shared standards had positive effect on interrater reliability and development of expertise around the assessment task. Other work that studied scorer discussions included that of Trace, Meier, Janssen (2016) who found that raters negotiating aspects of a rubric provided opportunities for the teachers to view the rubric differently, taking the perspectives of their co-scorers, which contributed to a deeper shared understanding of the rubric document. These studies support the notion that assessment processes and faculty conversation hold the true value and meaning of assessment work, beyond concepts of data reliability. It is the sharing of perspectives and articulating of "multiformity" (Broad, 2003), allowing for and hearing diversity within teaching and evaluation, that provides potential for individual and programmatic growth.

Influence of Faculty Emotions in Evaluative Tasks

Just as we understand and expect faculty members' individual pedagogical expertise and interactions within assessment groups to

influence their participation in assessment projects, internal emotional factors may also affect their judgments and evaluations of student writing. Work by Brackett, Floman, Ashton-James, Cherkasskiy, and Salovey (2013) explored the potential relationships between middle school teachers' emotions and the grades that they assigned student papers, contributing to an understanding of the way emotions may influence teachers' evaluations. They found that teachers were more likely to rate a paper positively after a priming task that involved positive emotions and negatively after priming task that involved negative emotions. The study results were in line with other work that suggested professional evaluations were influenced by assessors' emotional moods, from teaching evaluations (Floman, Hagelskamp, Brackett, & Rivers, 2016) to medical school entrance interviews (Redelmeier & Baxter, 2009). Evaluators' prior knowledge seems to also influence the degree to which they engage with writers and text of the papers that they are evaluating, personalizing the writers and ascribing a narrative to the writers' context and choices. Wiseman (2012) called this personalization of writers by evaluators "ego engagement" and explained that it may influence scoring and severity or leniency, as analyses suggested such patterns in holistic ratings in her study of scores assessing ESL writing proficiency exams.

College Instructors' Emotional Responses to Grading Student Papers

In a line of work exploring faculty emotional responses to their students' papers, recent survey research (Babb & Corbett, 2016) found college writing instructors reported strong emotional responses—most commonly disappointment, concern, and frustration—when they assigned failing grades to their students' papers. Babb and Corbett found that, in the case of instructors grading student papers, emotional feedback about their students' papers may serve as emotional "fuel for change" (Turner, 2010) in the sense that faculty can make adjustments to classroom instruction, assignments, pacing, or other elements of their course design in order to accommodate and answer the performance that they are seeing, positive or negative, in students' papers. Emotional responses were also reported in work by Caswell (2014), in which almost all faculty reported strong emotions while grading student papers. Think-aloud protocol analyses linked negative emotional responses to a violation of expectations; that is, during grading, faculty were frustrated when students did not meet the values that the faculty member had articulated in their assignments and instruction. These emotions, Caswell pointed out, provided feedback for the faculty members to adjust their own practices; thus, emotions, both positive and negative, played a critical role in the instructional relationship between faculty and students. Experiencing emotions while reading students' writing assignments may be particularly instructive for faculty as the emotions themselves serve as a type of feedback toward reflective practice (Edgington, 2005; Richardson, 1990).

Emotional Self-Regulation and Orthogonal Scoring

Given the influence of individuals' emotions and emotional states on their professional judgments, it follows that faculty members engaged in writing assessment experience emotional interference and biases during calibration discussions and individual paper scoring tasks. This may stem from individual moods (Bless & Fiedler, 2006) and the interaction of their highly personalized pedagogies (Kagan, 1992) with the papers they are reading, their interaction with one another, and the ecological context of the scoring situation (Dryer & Peckham, 2014), or a complex interaction of some or all of these variables.

As these factors play out within an assessment context, they likely have emotional consequences. Work by Thagard (2006) addressed the role of emotions in thinking and memory, building on a large body of work in psychology about "hot cognition" that spans over 50 years. By working to understand when emotions can promote clear thinking as opposed to when they get in the way of it, he called for "emotionally coherent decision making" (Thagard, 2006, p. 21) and an "emotional rationality" (Thagard, 2006, p. 11) in which individuals are able to acknowledge and integrate emotional intuition with calculated decision-making. Emotional intelligence theorists are similarly interested in individuals' abilities to perceive and regulate their emotions, with emotional intelligence understood as "an ability to recognize the meanings of emotions and their relationships and to use them as a basis in reasoning and problem solving…using emotions to enhance cognitive activities" (Mayer, Salovey, Caruso, & Sitarenios, 2001, p. 234). In this way, emotional intelligence and emotional rationality are understood not just as the ability to control emotions, but also to leverage them to accomplish a task.

As writing assessment directors work to promote meaningful experiences for faculty scorers, they also want the investment of time and resources into assessment projects to yield meaningful data to inform interventions and actions to improve student learning (Ikenberry & Kuh, 2015). Thus, it is important that scorers are able to assess each student paper independently from one another—that is, score one paper at a time—and to score each category of the rubric as relatively independent from the others. This can be especially difficult given the interconnected, systemic nature of language and argument and the levels involved in writing assessment tasks and the evaluative lenses of teachers as individuals, which can bring their own intricacies and layers, moods and (mis)understandings.

Even with these challenges and imprecisions of writing and its assessment, there is great value in the process itself, as it lends itself to important faculty conversations during evaluation and regarding actions to take around the assessment results. Thus, helping scorers to maintain orthogonality (Lomax, 2001), or the separateness of categories in a rubric, is important as it contributes to useful, actionable information in terms of programmatic performance (Kinzie, Hutchings & Jankowski, 2015). Statistically, orthogonal constructs are those which are "nonredundant and independent" (Lomax, 2001, p. 305). Applied to rubric-based scoring, this means that a score in one category does not influence a score in a different category in the mind of the assessor, or that the scores among

the rubric categories are not so highly correlated so as to suggest that all categories are assessing the same thing.

Locating and Valuing Emotions in Professional Tasks

Studies of faculty emotion during grading and assessment tasks represent movement of emotion away from "individualized, internally located, and privately experienced" (Stenberg, 2011, p. 349) phenomenon. Allowing space for discussion of emotions during professional tasks, or perhaps even foregrounding one's experience of them, recognizes their importance**.** Scholars such as Micciche (2007) and Ahmed (2013) have recently argued for emotions as valuable and, indeed, central to individuals' and groups' ways of making meaning in the world. Instead of looking suspiciously at our emotional responses—whether to student papers or colleagues' responses to them—rhetoric of emotion scholars would have us consider how emotion is "necessarily engaged in our classrooms" (Stenberg, 2011, p. 351) and other spaces.

Impetus and Context for Current Study

Given the complexity of the assessment task and the variability of experiences that teachers bring to bear on their assessment of student writing, I added an "overall response" category to the rubrics used in two recent assessment projects in an attempt to ease some difficult norming sessions in which some veteran teachers, new to assessment, were struggling with assessment scoring tasks. As the assessment facilitator for both of these projects, I added this category so the scorers had a place to locate and articulate their responses to a paper, particularly when they felt that other categories were not capturing their reactions to and evaluations of student writing. This new overall response category seemed to provide a useful talking point during calibration sessions and seemed to ease the scorers into the evaluation task at hand, allowing them a place to log what they "really thought" about a paper, in addition to the scoring that they did across the established rubric categories.

Once both of these assessment projects had ended, I wondered whether this additional category was worth retaining in future projects, given that the scoring rubrics were already rather long and the assessment work arduous. Specifically, I was interested in the function of this overall response category and set out to investigate it as a strategy to mitigate scorers' emotional responses to scoring and as a way to understand how scorers may self-regulate during assessment tasks. To this end, I designed this inquiry to address the following questions:

Research questions

1. Is the overall response category functioning, qualitatively and quantitatively, to help scorers maintain independence across categories?
2. How do faculty evaluators report using the overall response category as a means to self-regulate their scoring? To what extent, if any, is this category a useful tool for faculty scorers?
3. Statistically, how does this overall response category relate to the regular rubric categories? Which categories, if any, are significant predictors of this category? Are significant predictors of the overall response category similar across the different assessment projects?

Method

Data from this project came from two writing assessment projects at a public mid-sized university in the West. One of the assessment projects involved a core writing course from the University's Rhetoric and Writing Program, Composition II. The second project was part of the University's Writing across the Curriculum assessment, which pulled papers from a campus-wide writing portfolio. All data collection reported here was completed with approval from the University's Institutional Review Board.

Composition II Research Writing Course Assessment

Evaluators and project context. Scorers for the Composition II project were selected from faculty who applied to work as assessment fellows for a summer stipend (See Appendix A for Call for Applications). All five faculty assessment fellows had been teaching core writing courses for at least 10 semesters, and all faculty had at least eight semesters of experience teaching these courses at this university. As part of their summer assessment fellow contract, each evaluator was asked to attend eight hours of group norming and score about 60 student papers. They met several times after scoring data were available to plan faculty development around project results, thus working to "close the loop" on assessment by using findings to improve student learning.

Artifacts. Student papers from the one-semester research writing course, Composition II, required of all liberal arts, school of public affairs, education, and health science majors were submitted by course instructors for this assessment project. About 15% of the final papers from fall and spring semesters were randomly selected, resulting in 165 papers in the sample. These papers were then de-identified of student and instructor names in preparation for the evaluation.

Composition II was taught by a group of full-time instructors who use different topics to teach rhetoric-supported research writing based on stasis model, so the final paper for the course is a “stasis map,” which asks students to take a controversial issue through the stasis of definition, evaluation, cause, and/or proposal using library-based academic research. These stasis map papers ranged from eight to 12 pages long.

Scoring rubric. The Composition II scoring rubric, with detailed category and dimension descriptors at each of the four rating levels, had been in place for four summer assessment projects prior to this summer. It had strong rater reliability and consensus among program faculty that it represented the course values and goals.

Rubric categories. The 10 categories of this rubric remained unchanged for several years during its use. Several of the categories assess source use, data that are shared with the library, as librarians provide instructional support to this research-based course.

Table 1
Summary of Composition II Assessment Rubric and Reliability on a 4-Point Scale

Category	Description	Reliability*
Rubric Category		
Setting up the Argument	Writer's ability to create exigency, to acknowledge and situate the audience, and to establish an individual voice.	.37
Thesis	Writer's ability to make a clear arguable claim, though it may be delayed or implied.	.31
Argument Qualities	Explicit demonstration of theoretical elements. Writer is able to maintain rigorous analysis (as opposed to summary/exposition), to show awareness of multiple views and counter arguments (which are either conceded or rebutted), to avoid logical fallacies, and employ Toulmin and stasis argument principles and rhetorical appeals	.58
Development and Reasoning	Writer's ability to develop support for assertions-- evidence, specifics, reasoning, analysis, and definition-- throughout the paper. Writer has effectively joined an ongoing scholarly conversation. Idea development also shows focus, coherence, and a logical progression of ideas. Argument length is appropriate.	.62
Audience Awareness	Writer's ability to target a specific critical audience that is not the teacher or the general public.	.67
Organization	Writer's ability to control and purposefully guide the reader through the argument. The structure of the argument is logical and reader-based, and not structured around the sources. The argument exhibits effective transitions and topic sentences.	.46
Grammar, Mechanics, and Style	Writer's command of structure/syntax (including sentence boundaries involving comma splices, fused sentences, fragments), grammatical norms, punctuation, spelling, proofreading/editing, and diction.	.27
Sources:	Ability to execute chosen documentation style	.33
Technical Competence	(MLA, APA, etc.) including formatting (heading, pagination, paragraphing, margins, block quotes, etc.) and citations (in-text and works cited/reference page).	
Sources: Evaluation & Selection	Ability to deliberately select sources; evidence that library research tools were used to locate credible, relevant information, including scholarly sources (i.e. books, scholarly articles, government documents).	.46
Sources: Interpretation & Integration	Ability to analyze and interpret sources for critical audience; establish context for sources; use sources to justify argument; integrate and create balance between source voices and own voice.	.46
Added Category		
Overall response to paper	As a university faculty member, what is your overall response to the quality of this paper as a whole?	.67

Note. *Average difference between ratings. All papers were scored by two raters; some were scored by three.

Overall response. The "overall" category was not an average of other components. Rather, scorers were asked to rate the paper based on their response to it. I added this category in an effort to help faculty manage the scoring process and to provide a space for them to respond personally to the paper.

Campus-Wide Writing Portfolio Assessment

Artifacts. A total of 231 papers from undergraduate writing portfolios were used for this part of the project. The writing portfolio is a

graduation requirement of all students at this university, and students are advised to submit their portfolios of two papers after they have earned 90 credit hours and/or at least two semesters prior to graduation. The writing portfolio serves as an assessment of general education, and, with its newly revised rubric, the program also endeavors to provide meaningful data to programs and majors about their students' writing performance. In addition, it provides an intervention opportunity for students needing additional writing instruction before graduation.

Scoring rubric. Revised as part of a larger general education revision and implementation project, the scoring rubric for the writing portfolio had been in place for 6 months prior to this project. The rubric was created in collaboration with a faculty-led assessment committee, the Writing Across the Curriculum Director, and campus librarians, in consultation with the Association of American Colleges and Universities' *VALUE* rubrics about Information Literacy and Written Communication.

Table 2
Writing Across the Curriculum Writing Portfolio Assessment Rubric Categories and Descriptions

Category	Description
Rubric Category	
Purpose, Context, Claim, Perspective, Thesis, Hypothesis	The extent to which the writer presents a thesis statement, makes a primary claim, or clearly states the purpose of the essay/report. Degree to which this statement, or collection of statements, demonstrates a thorough understanding of the context and assigned task(s) and remains the focus of the writing throughout, and ideas/themes are fully identified and developed.
Development & Reasoning	The extent to which writer uses appropriate, relevant, and compelling content to illustrate mastery of the subject, conveying the writer's understanding, and shaping the whole work.
Critical Thinking	The extent to which the writer provides logical and specific details, appropriate for the discipline, to support claims. The degree to which, when appropriate, the writer thoughtfully considers multiple viewpoints. Conclusions are based upon presented evidence.
Genre & Disciplinary Conventions	The extent to which the writer demonstrates detailed attention to and successful execution of conventions particular to the specific discipline and/or writing task including content, presentation, formatting, and stylistic Choices.
Control of Syntax & Mechanics	The extent to which the writer uses language that skillfully communicates meaning to readers with clarity and fluency and is virtually error free.
Organization	The extent to which the writing demonstrates an effective pattern of organization consistent with its purpose. Paragraphs reflect appropriate level of thought and development. Paragraphs are effectively structured and ordered. Writer employs clear and appropriate transition.
Source Selection	The extent to which the writer demonstrates selection of high quality, credible, relevant sources to develop ideas that are appropriate for the discipline and genre of the writing.
Source Integration	The extent to which the writer integrates the claims and ideas of others with its own accurately and responsibly. The extent to which the writer uses sources effectively and integrates them smoothly, paraphrasing and occasionally directly quoting authorities to help substantiate or support its own point(s).

Source Use **Ethical Use of Sources**	The extent to which the writer uses sources ethically and responsibly, as demonstrated by correct use of citations and references. Uses all of the information use strategies: paraphrase, summary, and quoting in ways that appear true to original context.
Added Category	
Overall **Response to Paper**	As a university faculty member, what is your overall response to the quality of this paper as a whole?

Evaluator and project context. The writing portfolio office had six faculty members who assessed student portfolio submissions. For the purpose of the project reported here, analysis focused on a single scorer, the only one who did not have explicit instruction in writing pedagogy. I focused on a single scorer because, at this time, our budget and processes did not yet allow for regular double-scoring of student portfolios, so a measure of inter-rater reliability was not yet available. Writing portfolio assessment was not as developed as the assessment of this university's composition program, but plans were in place to hire and train additional faculty scorers, using data from this project to inform those norming and training sessions.

The scorer, Simon, had a Master's degree and had been teaching in the liberal arts for over 12 semesters. He had been scoring writing portfolios for three semesters. All of Simon's courses were designated Writing Intensive in a discipline known as offering writing-focused courses. For the purposes of this project, data from the three most recent months of Simon's writing portfolio scoring were used.

Understanding Scorers' Use of the New Rubric Category

In the qualitative part of this project, I wanted to understand how faculty scorers were using the overall response category in their scoring practice, whether they found it useful or burdensome, and what role, if any, it might play in managing the complex cognitive and emotional task of writing assessment. In particular, I was curious about whether they were using the new rubric category to manage their personalized responses to the papers, given what Kagan (1992) referred to as teachers' highly personalized pedagogies and how they might be using the category to manage the interplay of emotions and thought (Mayer et al., 2001; Thagard, 2006) while scoring.

Participants. The six scorers described above, five from the Composition II assessment project and one from the Writing Portfolio assessment, provided anonymous written responses on the scorer questionnaire.

Scorer questionnaire. This four-question form asked scorers to consider the way they use the overall category in scoring student artifacts, whether they found the category useful, and to envision whether their scoring might shift without the category. Items on the scorer questionnaire are listed in Appendix B. Participants hand-wrote their responses onto the questionnaire during a particularly intensive week of scoring activity, then returned their forms individually to the author at their convenience. Most returned their questionnaires the same day.

Results

First, I analyzed the scorer questionnaire to understand the faculty scorers' experience and use of the new category. Next, I turned to the quantitative data, the rubric category scores, to understand whether the overall category was functioning statistically to help the faculty scorers maintain independence across scoring categories.

Qualitative Scorer Questionnaire

The six participants wrote an aggregated 1,273 words in response to the four questionnaire items. I typed their responses into the qualitative software NVIVO to assist with the task of coding. I then analyzed the questionnaire data using grounded theory as it is described by Strauss and Corbin (1990) and elaborated in the methodological text by Charmaz (2009). This approach allowed for emergent themes in the data, as opposed to approaching analysis with *a priori* codes or set expectations.

During the coding processes, I followed the recommendation of grounded theorist Glasner (1998), who recommended use of gerunds as category labels, as they better capture the "sense of action and sequence" happening within the data (Charmaz, 2009). The coding process was highly recursive and involved checking back with the questionnaire data throughout the analyses, using the language of the participants as coding categories whenever possible. In so doing, I remained open to codes and categories

emerging from the data, grounded in participants' explanation of their experiences, which are key features of grounded theory data analyses (Strauss & Corbin, 1990). Once I had identified initial codes in the data via the open coding process, I "coded the codes," revisiting the questionnaire data. From this process, I arrived at the axial categories. The process is presented in Table 3, with the bottom row representing the early, open coding process and the middle row of the table representing the second phase of categorizing, axial coding, a stage that seeks dimensionality within each coding category (Strauss & Corbin, 1990). The top layer of the table represents the final categories. Table 3 is modeled after that of Anfara, Brown, and Mangione (2002).

Table 3
Coding Iteration Table: Scorer Questionnaire About 'Overall Response' Category

THIRD ITERATION: APPLICATION TO THE DATA SET				
Column A	**Column B**	**Column C**	**Column D**	**Column E**
Regulating emotions while scoring	Accounting for what rubric categories do not	Responding to paper emotionally	Managing the scoring task	Compartmental-izing
SECOND ITERATION: AXIAL CODE SUBCATEGORIES				
Managing frustration Moving on	Logging student effort Rewarding paper Punishing paper	Getting angry at paper Being impressed with paper	Relying on overall category as check	Maintaining Independence from other categories
FIRST ITERATION: INITIAL CODES				
Processing Matching feelings Subjective space Venting space Tension Handling issues with emotional responses	Giving credit Admiring paper Admiring effort Not concrete; abstraction	Holding place for emotion Score feeling right Acknowledging scorers' own reaction	Check and balance Self-regulate scoring Dependent on other categories Gut checking	Triggers sustainability Allows moving on to next paper Helping with fatigue
DATA	DATA	DATA	DATA	DATA

The sections that follow enumerate the third iteration, in the top row of Table 3, by providing examples and explanation of these categories from the questionnaires.

Regulating Emotions While Scoring

The code of regulating emotions while scoring refers to the ways that scorers used the overall category as a specific place to log or express emotions, in an explicit effort to keep these feelings from impacting other aspects of their scoring.

Managing frustration. Two of the scorers reported using the overall category to house upsetting emotions such as anger and frustration. As one scorer explained, "I know it [overall score] doesn't count in the same way [toward our assessment], so if I'm angry or upset with a paper or see that a student has put in effort then that's a place where I can let that out without it counting." Another wrote, "If I'm pissed off or angry I can put that here." This colleague echoed this idea of the overall score as a place to house negative emotions, explaining that the category, "allows for me to log tension that arises from doing assessment."

Moving on. The new category appeared to serve not just as a holding place for negative emotions, but also as a means for scorers to address the task at hand: that of scoring efficiently and moving through assessment of papers. About this, a faculty member explained that the category is useful because it "allows me to score the next paper with a cleaner palette." This scorer's colleague expressed similar sentiment and explained their use of the overall category, explaining, "If things about the paper appealed to me or bothered me logically, professionally, or emotionally, I will weigh them heavily in the overall category and move on." Yet another seemed to directly address the idea of moving forward with the assessment task at hand, as they wrote about their use of the overall category, "I believe this emotional response allows and triggers sustainability," suggesting use of the category helped this scorer to move on to read, and assess, the next student paper.

Accounting for What Rubric Categories Do Not

Five of the six scorers explained using the overall category as something of a catch-all box in their assessment, tracking features of papers they thought were not accounted for in other aspects of the scoring rubric. They reported using the category to laud or denounce papers about which they had strong emotional reactions. For instance, one scorer explained that the category provided "room to reward effort and taking on a daunting task," which was an interesting comment given that the students will not see their scores on the evaluation. Another explained, "It gives me a place to say whether the paper represents good work from our program."

Other faculty assessors also reported using the overall category to recognize a paper whose excellence may not be accounted for via the rubric. For instance, one scorer explained that the category "allows me to give credit for what works well even when it isn't clearly represented in the other criteria-- sometimes the total is greater than the sum of its parts, and this category takes that into account." Another scorer also reported reading beyond the page itself and into the context when they used the category, "I think about 'is the student limited by the assignment? How difficult is the task? Is there some place I can't help and let the assessment know I find this unfair?'" In these ways, the faculty scorers reported using the new category to capture the unseen and their instincts regarding a student's circumstances around writing a paper.

Compartmentalizing

Many of these same scorers also reported using the overall category when they are either frustrated or enamored with a paper, but reported doing so as a way to independently score the other rubric categories. During norming sessions, scorers were introduced to the concept of orthogonality (Lomax, 2001) with regard to scoring, and they were asked to assess each rubric category independently of others, to the extent that they were able to do so given the complex system of writing they were evaluating. Thus, several of the scorers used the overall category as a way to help maintain a more compartmentalized view of the paper and the individual categories within the rubric. One scorer explicitly addressed this on the questionnaire when asked about how they used the overall category: "It has been helpful to use as a place to acknowledge my overall feeling or reaction to a particular paper, without affecting an unrelated criterion."

When asked about whether their scoring would be impacted if the overall category were eliminated (Appendix B Question 3), fellow scorers echoed the usefulness of the category in their questionnaire responses. One explained, "If I didn't have it, my gut reaction would find other ways of coming out so that the papers I hate would be scored lower than they technically deserve and vice versa." Another responded, "If we didn't have the [overall] category, I imagine I might find somewhere else to be more generous or harsh, though, out of a desire to give credit or vent frustration," which suggested the category was working to promote orthogonality.

Managing the Scoring Task

Within the data set of questionnaire responses, this code referred to ways that scorers used the holistic category to accomplish their work of efficiently and thoughtfully assessing student papers. For many, they had developed strategies of using the overall response category as a means to compare to their rubric scores before moving on to assess the next artifact. In this way, the overall score had become a reflective moment during scoring.

Many of the faculty scorers reported using the overall rubric category as a means of verifying and auditing their scoring across the other rubric categories, a final step as they assess an individual paper. So, they explained that they assign an overall response score last in their assessment of a paper. As one faculty member explained, "Unlike other categories, it [overall score] forms amorphously over the course of assessment, only coming fully into view as I finish my reading; however, unlike the other categories, there often is not something concrete that I can put my finger on for this category." So, for this reader, assessment of this category functioned as a gestalt, holistic score of the paper. In terms of process, other scorers also reported scoring the category at the end of the paper assessment, "I wait to score it last and start with my 'gut instinct,' then review the rest of the scores to see if it is consistent with them-- although that review rarely changes the overall score."

Other scorers, however, reported a more recursive relationship between the overall category and the rest of the rubric and explained, "If the [overall] score does not 'feel' right, I will review the paper against the rubric and my scores to make sure I scored it appropriately." This scorer's colleague explained similar corroborative use of the category, "Sometimes if the overall score isn't consistent with my category scores, then I'll revisit my category scores. So it can work for me as my own checks and balances system."

Quantitative Relationships Among Rubric Categories

To better understand the quantitative relationships among the conventional and overall categories of each scoring rubric, I correlated the rubric categories of the papers, then ran linear regression analysis predicting the overall category from the rubric categories for the Composition II project and the Writing Portfolio paper scoring.

Composition II: Relationships Between Categories and Predicting "Overall" From 10 Conventional Categories

Correlations between rubric categories were calculated for the 165 papers, which are reported in Table 4. All rubric categories were significantly correlated with one another and with the overall score, with correlation coefficients ranging from .44 - .83. Correlations between rubric categories and the overall scores ranged from .54 - .83. The moderate to strong, but significant, correlations between rubric category scores were expected given the interconnected nature of writing and writing assessment. For instance, the elements of argument qualities, development and reasoning, and audience awareness were highly correlated (correlation coefficients around .77). Examining the rubric (Table 1), it makes sense that these writing features would be related to one another. Similarly, features of source quality and integration were correlated at a level of .68. Notably, the largest correlation coefficients in the overall rubric category were also significant predictors in the regression model reported below, which is to be expected (Bobko, 2001).

Table 4
*Composition II Correlation Coefficients Between Rubric Categories**

	1	2	3	4	5	6	7	8	9	10	11
1.Setting up Argument	-										
2. Thesis	.67	-									
3. Argument Qualities	.70	.74	-								
4. Develop & Reasoning	.63	.67	.77	-							
5. Audience Awareness	.63	.66	.77	.74	-						
6. Organ-ization	.57	.53	.64	.67	.68	-					
7. Grammar Mech Style	.48	.50	.53	.53	.57	.50	-				
8. Sources: Tech Comp	.44	.46	.57	.53	.54	.51	.44	-			
9. Sources: Eval & Select	.50	.54	.62	.68	.67	.53	.46	.63	-		
10. Sources: Intp & Integ	.63	.59	.74	.76	.68	.57	.46	.63	.68	-	
11. Overall	.69	.70	.83	.82	.79	.67	.54	.64	.74	.81	-

Note. *All correlations significant at the .01 level.

A simple linear regression was used to test whether the 10 conventional categories were predictive of evaluators' ratings of the overall score category, presented in Table 5. The results of the regression indicated that the conventional categories together explained 82% of the variance in the overall scores (R^2 = .83, F(10, 140) = 70.20, p < .01). In this model, four categories significantly contributed to the overall score category. These significant predictors included "Argument qualities" (β = .21, p < .01), "Development and reasoning" (β = .18, p < .01), "Source evaluation and selection" (β = .15, p < .05), and "Source interpretation and integration" (β = -.19, p < .01).

Collinearity occurs when two independent variables are so closely related that they both behave similarly and powerfully in a regression equation (Miles & Shevlin, 2001). The collinearity statistics for the regression model are presented in Table 5 in order to explore whether any of the rubric categories were statistically so similar as to suggest that collinearity was a problem. Given the nature of the measurement task here, using a rubric to assess a student paper, we can expect the rubric scores to be highly related due to the systemic nature of writing. However, the assessment task at hand called for scorers to evaluate each paper for the rubric categories independently, maintaining as much distinction as possible between categories.

As indicated in Table 4, rubric categories were significantly correlated. With regard to the collinearity statistics in Table 5 the

Variance Inflation Factor (VIF), a measure of collinearity tolerance, was below 4, the recommended cutoff for this statistic before error in the regression equation due to collinearity becomes a concern (Miles & Shelvin, 2001). Argument qualities, the only category for which the VIF was over the recommended 4, was only slightly above it at 4.25. The 'tolerance' column of Table 5, presents values between 1 - 0, with those closer to 0 indicating problems of collinearity.

Table 5
Composition II Linear Regression Predicting "Overall Response" from Rubric Categories

Rubric Category	Overall Response		Collinearity Statistics	
	Standardized Beta	Significance	Tolerance	VIF
Setting up the Argument	.06	.23	.42	2.38
Thesis	.05	.39	.38	2.62
Argument Qualities	**.21**	**.00***	.23	4.25
Development and Reasoning	**.18**	**.01***	.26	3.80
Audience Awareness	.12	.07	.28	3.51
Organization	.05	.31	.45	2.24
Grammar, Mechanics, and Style	.01	.91	.61	1.65
Sources: Tech Competence	.06	.21	.47	2.12
Sources: Evaluation and Selection	**.15**	**.01***	.37	2.68
Sources: Interpretation and Integration	**.19**	**.00***	.30	3.35

Note. * p<.05
** p< .01

Campus-Wide Writing Across the Curriculum Portfolio Assessment

I then conducted the same set of analyses on the Writing Portfolio data. First, I calculated correlations between the rubric categories for 231 writing portfolio papers that Simon had scored across three consecutive months. These are reported in Table 6. As with the Composition II correlation calculations, almost all of the correlations were significant at the moderate to high range. Categories involving source use, which included sources selection, integration, ethical use, had the highest correlations (see Table 2 for descriptions of these categories). With regard to the overall category, almost all of the other rubric categories were correlated at a moderate level (.28 - .55). When considered as an overall model, in the regression reported below, only three were significant predictors, as explained in the section that follows.

Table 6
*Correlations Between Rubric Categories for Writing Portfolio Rubric**

	1	2	3	4	5	6	7	8	9	10
1. Purpose; Context; Thesis	-									
2.Development & Reason	.50	-								
3. Critical Thinking	.49	.44	-							
4. Genre & Disc Convent	.47	.48	.40	-						
5. Syntax & Mechanics	.33	.37	.28	.35	-					
6. Organization	.58	.50	.39	.42	.38	-				
7. Source Select	.21	.30	.20	.23	(.12)	.28	-			
8. Source Integration	.31	.38	.28	.36	.18	.33	.77	-		
9. Source Ethical Use	.31	.39	.29	.31	.20	.37	.78	.81	-	
10. Overall	.48	.55	.43	.42	.28	.44	.29	.35	.35	-

Note. *All correlations significant at the .01 level except where ( )

Then, as with the Composition II sample, I ran a simple linear regression using the rubric associated with the writing portfolio. The regression tested whether the nine conventional categories were predictive of the evaluator, Simon's, rating of the overall response category.

Table 7
Writing Portfolio Linear Regression Predicting "Overall Response" from Rubric Categories

Conventional Category	**Overall Response**		**Collinearity Statistics**	
	Standardized Beta	**Significance**	**Tolerance**	**VIF**
Purpose, Context, Claim, Perspective, Thesis, Hypothesis	**.16**	**.03***	.52	1.92
Development & Reasoning	**.31**	**.00***	.57	1.76
Critical Thinking	.10	.12	.68	1.47
Genre & Disciplinary Conventions	**.14**	**.04***	.65	1.54
Control of Syntax & Mechanics	-.03	.64	.79	1.26
Organization	.09	.21	.56	1.78
Source Use: Source Selection	.04	.63	.35	2.86
Source Use: Source Integration	-.01	.93	.27	3.67
Source Use: Ethical Use of Sources	.05	.63	.27	3.63

Note. * p<.05

The results of the regression indicated that the conventional categories together explained 39% of the variance in the scores (R^2 = .42, F(9, 202) = 16.07, p < .01). In this model, three categories significantly contributed to the overall response category. These significant predictors included "Purpose" (β = .20, p < .05), "Development" (β = .31, p < .01), and "Genre" (β = .14, p < .05). Table 7

presents data from the categories in this model. With regard to collinearity, all VIFs were under the recommended cutoff of 4, and the tolerance values were well above 0 in the 0 - 1 range (Miles & Shelvin, 2001).

Discussion and Implications

Findings from this exploratory study suggest that the inclusion of an overall response rubric category was a useful time investment for the faculty conducting the scoring in this project. Although it added an extra category to already-long scoring rubrics, it also provided a point-of-reference for faculty to gauge their "gut" responses to papers and house their emotional reactions to the student artifacts they were scoring, as reported in the qualitative data. The quantitative data suggested scorers who used this additional category in the assessment projects were scoring independently, thus adding confirmation to their reports that the overall category helped them to compartmentalize and assess each aspect of the rubric individually.

Scorers Maintaining Independence Across Categories

Across both of these assessment projects, the rubric predictors of the overall score were not so highly related as to indicate they were measuring the same or overlapping phenomena (Bobko, 2001). This determination of collinearity was made by reviewing the correlation coefficients, reported in Tables 4 and 6, and the collinearity statistics, reported in Tables 5 and 7. It is also important to consider the measurement context, in this case the assessment of student papers via rubrics; we expect that features of student papers will correlate moderately.

Differences Between Composition II and Writing Across the Curriculum Assessment Projects

The different assessment projects had different rubrics, scorers, and data sets, as reported above. There were some similarities between the assessment rubrics when comparing them in Tables 1 and 2. However, the regression equations revealed differences in terms of the weight of the predictors (Tables 5 and 7) and which predictors were significant in terms of their contribution to the overall score. Reasons for this may include the different assessment ecologies of the two projects (Dryer & Peckham, 2014). In the case of the Composition II project, the faculty scorers were all instructors of the course, familiar with the assignment, and the student papers were assignments that were relatively homogeneous—they were stasis map assignments from various sections of the same course. Where there was some variety in terms of each section instructor's interpretation of the assignment, as allowed by the program curriculum, the assignment had common features and outcomes.

In contrast, student papers from the writing portfolio's writing across the curriculum assessment ranged across disciplines and levels (upper and lower division courses), running the gamut from mechanical engineering lab reports about fluid dynamics, to nursing case studies about intersections between mental health and surgery recovery, to criminal justice reaction papers about corporate violence, and a broad range of disciplines in between. Heterogeneity within the sample may have contributed to weaker correlations among the rubric categories of the writing portfolio project, even though the correlation coefficients for each project were significant. Finally, recall that the writing portfolio data include a single scorer, whereas the Composition II scoring data represent evaluations from five faculty scorers.

The Overall Response Category as Self-Regulation and Reflection Strategy

Analyses of the faculty scorers' questionnaire responses suggest that the new category helped scorers leverage their emotional responses to facilitate the task of assessment scoring. In this way, the overall response rubric category was contributing to emotional self-regulation (Mayer et al., 2001) and perhaps to more emotionally intelligent scoring. Scorers reported using the new rubric category to manage their particularly strong responses to student papers, including emotions that ran both highly negative and highly positive. Further, and perhaps most importantly, they also reported leveraging this rubric category to manage the complex task of scoring the papers to the rubric, suggesting that the category helped them adopt a balance between emotions and cognition, contributing to what Thagard (2006) referred to as "emotional rationality" in which emotions are not just managed and controlled but are capitalized upon to help with difficult decisions.

The reported helpfulness of this category seems particularly important in light of earlier work by Caswell (2014) and Edgington (2005), whose findings suggested faculty members' emotional responses during grading provided feedback that served as motivation toward their reflection and instructional change. Assessment contexts are different from grading contexts, but both involve responding to student writing, and, thus, faculty may experience strong emotional responses. Faculty assessment scorers may have different emotional experiences because their roles do not call for direct instructional intervention; unlike reading their own students' papers, they cannot make changes to their course context the next day based on what they read in the papers. Future work might track the ways in which faculty scorers' emotional responses to student papers influence their own instructional practice, even in cases when the papers are not from faculty members' students or courses.

Appropriating Assessment Categories for "Internal Use Only"

With regard to yielding actionable programmatic assessment data, holistic, impressionistic scoring is controversial for reasons of reliability (Charney, 1984; Huot, 1990) and reductivity (Hamp-Lyons, 2016; Neal, 2011). That is, holistic scoring is often difficult to calibrate among groups of raters and then problematic with regards to suggesting action for change to improve teaching and student learning. While the ultimate usefulness and reliability of holistic scoring for programmatic assessment is tricky, the practice of scoring holistically may be a useful tool for raters to regulate their own emotional responses to the writing that they are asked to assess; it may serve as a "gut check" of their initial responses, which they can then learn to unpack and conduct their evaluation work, either via group discussions or personal reflection.

The recognition of multidimensionality of writing and use of rubrics and other tools have moved us ahead in terms of meaningful assessment data and meaningful conversations about teaching and learning (Adler-Kassner & O'Neil, 2010). With that said, impressionistic scoring may be a useful tool for facilitators to help raters name their emotional responses to the writing that they are asked to assess, monitor their initial responses, and learn to address those responses and conduct their assessment work. As such, impressionistic scoring may be viewed as a strategy for learning to use a rubric and assess multiple dimensions of a piece of writing and a way for faculty to gain further insight and reflect on their own values (Broad, 2003).

Rubrics as Evolving Process Documents Supporting Faculty Thinking

Rubrics, like any tool, can be wielded and misused; they allow for quantification and reduction of complex phenomena regarding learning and communication, which adds to the potential for misuse (Broad, 2009; Hamp-Lyons, 2016). However, rubrics can also provide a means to articulate values and make programmatic comparisons chronologically and among groups. As rubrics are modified across time and location, we should consider categories that facilitate their use, in addition to those that directly evaluate student learning. Gut categories, such as those implemented in the assessment projects detailed here, may serve a facilitative role. Supportive rubric categories may serve as personal spaces where scorers may reflect on their own practice as evaluators and teachers (Adler-Kassner & O'Neil, 2010), supplementing professional conversations among groups of scorers (Colombini & McBride, 2012) or scoring pairs (Jølle, 2014). Further, while the overall category reported here was both impressionistic and holistic, and not likely to yield actionable assessment data, such categories may facilitate scorers' discussions of the work they are evaluating. Thus, they have the potential to reveal misunderstandings, assumptions, and complexities inherent in evaluating writing. The more we can identify strategies and means to engage faculty in the assessment process itself—which is best practice—the easier it may also be to arrive at the important "next steps" and "closing the loop" of actions to take regarding programmatic assessment data (Ikenberry & Kuh, 2015).

Limitations, Implications, and Future Directions

As writing assessment facilitators strive to make assessment work more meaningful and accessible for faculty participants, exploring the intersection of faculty members personalized pedagogies (Kagan, 1992), beliefs about student learning and writing (Neely, 2017), and emotional responses to evaluating student writing (Brackett et al., 2013), they may also want to identify strategies and tools to make assessment scoring more easeful. Additional data, beyond that collected about the six scorers via the questionnaire in the current exploratory study, are necessary to better understand this phenomenon. Thus, accessing this interplay of factors further, via think aloud protocols of scoring and direct recording of scoring calibration sessions, as by Gebril and Plakans (2014), may help us to better understand how emotional responses play out, not only in particular rubric categories, but across all aspects of the rubric and scorers' assessment experiences. In addition, study designs that compare the overall response category, and other scoring strategies, on category independence may also be warranted. Identifying ways to capture, recognize, and leverage teachers' expertise (Osborne & Walker, 2014) via interviews and reflective journaling during assessment projects may also supplement the questionnaire-style data collected in the project reported here.

The current study advocates "capturing" scorers' emotional responses in a holistic category in order to help them regulate emotional responses while scoring students papers. Such an approach leans toward privileging rational efficiency (Stenberg, 2011) and a positivist approach that, perhaps while not dismissive of emotion, may seem suspicious of its value (Spelman, 1989). Future work might explore think-aloud protocols of scorers using an "emotional response" category in assessment in order to better understand the way emotions inform and assist faculty decisions during programmatic assessment tasks. In so doing, we might ask that faculty "stay with emotions" (Micciche, 2016) as a means to understand how their emotions help them create individual and collective meaning within evaluation contexts. Such an exploration would support the fundamental purpose and value of programmatic assessment: to understand and improve student learning.

Further exploring the role of faculty scorers' emotions in assessment contexts may include collecting data from formal and informal assessment conversations, individual scoring tasks, and faculty members' experiences after assessment, as the assessment project becomes a memory for faculty, which they take forward into their other professional practices, including teaching and future program evaluation.

References

Adler-Kassner, L., & O'Neill, P. (2010). *Reframing writing assessment to improve teaching and learning*. Logan, UT: Utah State University Press.

Ahmed, S. (2013). *The cultural politics of emotion*. New York: Routledge.

Anfara Jr, V. A., Brown, K. M., & Mangione, T. L. (2002). Qualitative analysis on stage: Making the research process more public. *Educational Researcher, 31*(7), 28-38.

Babb, J., & Corbett, S. J. (2016). From zero to sixty: A survey of college writing teachers' grading practices and the affect of failed performance. In *Composition Forum*, *34*(Summer).

Bless, H., & Fiedler, K. (2006). Mood and the regulation of information processing and behavior. In J. P. Forgas (Ed.), *Affect in Social Thinking and Behavior* (pp. 65-84). New York: Psychology Press.

Bobko, P. (2001). *Correlation and regression: Applications for industrial organizational psychology and management* (2nd ed.). Thousand Oaks, CA: Sage.

Brackett, M. A., Floman, J. L., Ashton-James, C., Cherkasskiy, L., & Salovey, P. (2013). The influence of teacher emotion on grading practices: A preliminary look at the evaluation of student writing. *Teachers and Teaching*, *19*(6), 634-646.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. University Press of Colorado.

Broad, B. (2009). Organic matters: In praise of locally grown writing assessment. University Press of Colorado. In Broad B, Adler-Kassner L, Alford B, Detweiler J, Estrem H, Harrington S, McBride M, Stalions E, Weeden S (Ed). *Organic writing assessment: Dynamic criteria mapping in action* (pp. 1-13). University Press of Colorado.

Caswell, N. (2014). Dynamic patterns: Emotional episodes within teachers' response practices. *Journal of Writing Assessment, 7*(1).

Charmaz, K. (2009). *Constructing grounded theory: A practical guide through qualitative analysis*. London: Sage.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*, 65-81.

Colombini, C. B., & McBride, M. (2012). “Storming and norming”: Exploring the value of group development models in addressing conflict in communal writing assessment. *Assessing Writing*, *17*(4), 191-207.

Condon, W., Iverson, E. R., Manduca, C. A., Rutz, C., & Willett, G. (2016). *Faculty development and student learning: Assessing the connections*. Indiana University Press.

Dryer, D. B., & Peckham, I. (2014). Social contexts of writing assessment: Toward an ecological *Construct* of the rater. *WPA: Writing Program Administration-Journal of the Council of Writing Program Administrators*, *38*(1), 12-41.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Edgington, A. (2005). What are you thinking?”: Understanding teacher reading and response through a protocol analysis study. *Journal of Writing Assessment*, *2*(2), 125-147.

Fives, H., & Buehl, M. M. (2012). Spring cleaning for the “messy” construct of teachers’ beliefs: What are they? Which have been examined? What can they tell us. *Educational Psychology Handbook, 2,* 471-499.

Floman, J. L., Hagelskamp, C., Brackett, M. A., & Rivers, S. E. (2016). Emotional bias in classroom observations within-rater positive emotion predicts favorable assessments of classroom quality. *Journal of Psychoeducational Assessment*. doi:10.1177/0734282916629595

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated

tasks. *Assessing writing*, *21*, 56-73.

Glaser, B. G., & Kaplan, W. D. (Eds.) (1998). *Gerund grounded theory: The basic social process dissertation.* Mill Valley, CA: Sociology Press.

Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching*, *8*(3), 381-391.

Hamp-Lyons, L. (2016). Farewell to holistic scoring. Part Two: Why build a house with only one brick? *Assessing Writing*, *29*, A1-A5. doi:10.1016/j.asw.2016.06.006

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

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201-213.

Huot, B. (1997). Beyond accountability: Reading with faculty as partners across the disciplines. In K. B. Yancey & B. Huot (Eds.), *Assessing writing across the curriculum: Diverse approaches and practices* (pp. 69-78). Greenwich, CT: Ablex.

Huot, B. (2002). Toward a new discourse of assessment for the college writing classroom. *College English*, *65*(2), 163-180.

Ikenberry, S. O., & Kuh, G. D. (2015). From compliance to ownership: Why and how colleges and universities assess student learning. In G. D. Kuh et al. (Eds.)*, Using evidence of student learning to improve higher education* (pp. 1-23). San Francisco, CA: Jossey-Bass.

Jølle, L. (2014). Pair assessment of pupil writing: A dialogic approach for studying the development of rater competence. *Assessing Writing*, *20*, 37-52.

Kagan, D. M. (1992). Implication of research on teacher belief. *Educational Psychologist*, *27*(1), 65-90.

Kinzie, J., Hutchings, P., & Jankowski, N.A. (2015). Fostering greater use of assessment results: Principles for effective practice. In G. D. Kuh et al. (Eds.), *Using evidence of student learning to improve higher education* (pp. 51-72). San Francisco, CA: Jossey-Bass

Kuh, G. D., & Hutchings, P. (2015). Assessment and initiative fatigue: Keeping the focus on learning. In G. D. Kuh et al. (Eds.), *Using evidence of student learning to improve higher education* (pp. 183-200). San Francisco, CA: Jossey-Bass.

Lomax, R. G. (2001). *An introduction to statistical concepts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion, 1*(3), 232-342.

McConnell, K. D., & Rhodes, T. L. (2017). *On solid ground: VALUE report 2017.* Retrieved from http://www.aacu.org/sites/default/files/files/FINALFORPUBLICATIONRELEASEONSOLIDGROUND.pdf

Micciche, L. R. (2007). *Doing emotion: Rhetoric, writing, teaching*. New Hampshire: Boynton/Cook.

Micciche, L. R. (2016). Staying with emotion. *Composition Forum, 34*(Summer). Miles, J., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. London: Sage.

Neal, Michael R. *Writing assessment and the revolution in digital texts and technologies*. Teachers College Press, 2011.

Neely, M. E. (2017). Faculty Beliefs in Successful Writing Fellow Partnerships: How Do Faculty Understand Teaching, Learning, and Writing?. *Across the Disciplines*, *14*(2).

Osborne, J., & Walker, P. (2014). Just ask teachers: Building expertise, trusting subjectivity, and valuing difference in writing assessment. *Assessing Writing*, *22*, 33-47.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62(3),* 307-332

Redelmeier, D. A., & Baxter, S. D. (2009). Rainy weather and medical school admission interviews. *Canadian Medical Association Journal*, *181*(12), 933-933.

Richardson, V. (1990). Significant and worthwhile change in teaching practice. *Educational Researcher*, *19*(7), 10-18.

Spelman, E. (1989). Anger and insubordination. In A. Garry & M. Persall (Eds.), *Women, knowledge, and reality: Explorations in feminist philosophy* (pp. 263-274). Boston: Unwin Hyman.

Stenberg, S. (2011). Teaching and (re) learning the rhetoric of emotion. *Pedagogy*, *11*(2), 349-369.

Strauss, A., & Corbin. J. (1990). *Basics of qualitative research*. Newbury Park, CA: Sage.

Strømsø, H. I., & Bråten, I. (2011). Personal epistemology in higher education: Teachers' beliefs and the role of faculty training programs. In J. Brownlee, G. Schraw, & D. Berthelsen (Eds.), *Personal epistemology and teacher education* (pp. 54-67). New York: Routledge.

Thagard, P. (2006) *Hot though: Mechanisms and application of emotional cognition.* Cambridge, MA: Bradford.

Trace, J., Meier, V., & Janssen, G. (2016). "I can see that": Developing shared rubric category interpretations through score negotiation. *Assessing Writing*, *30*, 32-43.

Turner, J. H. (2010). The stratification of emotions: Some preliminary generalizations. *Sociological Inquiry*, 80(2), 168-199.

Walvoord, B. E. (1997). From conduit to customer: The role of WAC faculty in WAC assessment. In K. B. Yancey & B. Huot (Eds.), *Assessing writing across the curriculum: Diverse approaches and practices* (pp. 69-78). Greenwich, CT: Ablex.

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*(3), 150-173.

Appendix A

Call for Applications

Writing Program Assessment Fellows

Call for assessment fellows:

We are looking for faculty to apply for positions as assessment fellows to participate in program-wide readings to be held during this summer. Please apply by sending an email to the Assessment Coordinator by April 24 to express interest and availability.

Application:

Include a brief letter of application answering the following questions:

1. Why does assessment interest you? Please describe why you would like to participate in this project.
2. What strengths will you bring to assessment work? Please describe the qualities and experience you possess that will allow you to work effectively on this collaborative project.

Expectations:

- All readers will be available for May planning meeting.
- Readers will be available for two 4-hour norming meetings in July for Composition II scoring.
- Interested readers will participate in October professional development session.

Appendix B

Questionnaire About Scoring Categories

1. As a scorer, do you find the “overall response” category useful? Why or why not?
2. How and when do you decide how you will score the “overall response” category? If there is a thought process, please describe it:
3. If we omitted the “overall response” category from the rubric, do you think it would impact your scoring of the other categories? Why or why not?
4. Would you say that you have an emotional response felt sense to scoring this “overall” category? That is, do you find yourself responding to this category differently compared to the other ones? Please explain.