Research Article

Contemporary Peer Review: Construct Modeling, Measurement Foundations, and the Future of Digital Learning

Ashley N. Reese, *University of South Florida*

Rajeev R. Rachamalla, *University of South Florida*

Alex Rudniy, *University of Scranton*

Laura L. Aull, *Wake Forest University*

David Eubanks, *Furman University*

Structured Abstract

- **Background:** In this article, we offer a study of peer review in a digital learning environment. Our analysis focuses on intrapersonal and interpersonal domains of the writing construct as they are enacted in the peer review process in terms of self-reflection and transaction. Our study is organized as a demonstration of the force of construct articulation, the usefulness of fairness as an integrative measurement framework, and the affordances of research in digital ecologies. Based on findings from our National Science Foundation funded research, we conclude with considerations for future peer review research.
- **Literature Review:** Despite the fact that most first-year composition programs utilize peer review, there is little writing studies research surrounding the practice of peer review (Haswell, 2005, p. 211). The studies that have addressed peer review generally find that peer review leads to

positive outcomes (Moxley & Eubanks, 2015; Ross, Liberman, Ngo, & LeGrand, 2016). Notably, peer reviews appear to help both the reviewer and the reviewee (Dochy, Segers, & Sluijsmans, 1999). Our understanding of the revision process is rooted in Flower and Hayes' (1981) social cognitive theory of writing. Combined with a need to expand models of the writing construct based on cognitive, interpersonal, and intrapersonal demands, our research seeks to fill the gap in acknowledging that the metacognitive nature of peer review is part of the construct of writing.

- **Research Questions:** Our research questions divide into three categories: the intrapersonal and interpersonal domain, forms of evidence, and digital learning affordances. We inquire into (1) the tone and quality of student self-reflection, as well as (2) the quality and tone of the peer review transaction. In the study of fairness evidence, we ask (3) what may be learned by investigating responses when student sub-groups are disaggregated according to gender, ethnicity, race, and English language learning. In the study of reliability evidence, we ask (4) what forms of evidence related to response consistency are useful in the analysis of peer review. In the study of validity evidence, we ask (5) how a precise definition of the writing construct lends precision to construct-related evidence. In terms of digital learning, we ask (6) what is the instrumental value of questions 1-5 in terms of demonstration of affordances to participate in the *MyReviewers (MyR)* peer review process.
- **Research Methodology:** Our research utilizes a sample of 837 students enrolled in first-year composition at a public research university, in particular their self-reflection ratings and the transaction ratings. These surveys were conducted voluntarily, presented to the students upon completing peer review (for reviewers) and the revision plan (for reviewees) as part of the *MyR* software.
- **Results:** The study shows that while self-reflection and transaction surveys received high to neutral ratings for helpfulness, politeness, and kindness, encouragement received only high or low ratings. In terms of fairness evidence for self-reflection, women believed their feedback was more polite and helpful. Similarly, Hispanic students believed their reviews were more helpful than non-Hispanic students did, and students who claimed proficiency in two or more languages felt their own reviews were more helpful than English-only speakers did. For fairness evidence for transactions, men were perceived as more encouraging than women in their feedback, while Hispanic students' reviews were no more helpful than non-Hispanic students' reviews. No statistically significant differences were found amongst English language learners and native English speakers. In relation to reliability evidence for

self-reflection, for the most part reliability reaches statistically significant levels. In terms of digital learning affordances, students in groups typically associated with low writing performance thrived in the digital learning platform when the construct included domains beyond the cognitive.

- **Discussion:** Based on these findings, there are three areas of consideration worthy of extended pursuit: 1) consider the advantages of expanded notions of the writing construct; 2) consider information analysis in terms of opportunity to learn; and 3) consider digital ecologies as a way to advance writing instruction for all students.
- **Conclusions:** This study provides unique insight that writing program administrators (WPAs) might utilize to inform their programs. A natural next step to implement the findings of our study would be for WPAs to systematically examine how evidence reacted to gender, ethnicity, and race is manifested within the classroom at their own institutions.

Keywords: corpus linguistics, first-year composition, peer review, student writing, writing analytics

1.0 Background

In 1984, Martin Nystrand was deeply involved in a metacognition study of student composing processes. Funded by the National Institute of Education, Nystrand focused his final report on the effectiveness of peer review in expository writing instruction. His study site was the University of Wisconsin at Madison where first-year students met weekly to review each other's writing in a studio environment. Nystrand used a mixed method empirical approach. Qualitatively, he videotaped students working in groups to investigate interactions and composing process awareness. Quantitatively, he used a quasi-experimental (non-random) paired group design to study score gains on writing performance. Nystrand found no statistically significant differences between studio and traditional groups at the beginning of the semester ($F = .079$; $p < .05$); however, by the end of the semester, the scores of studio students were greater than those of the non-studio students ($F = 3.018$; $p < .001$).

Based on these group differences, Nystrand then provided a series of reasons explaining the relationship between score gains and peer review based on survey results: Studio students treated revision as a reconceptualization, not as simply editing; they viewed their readers as collaborators in writing improvement, not as external judges; they gave additional emphasis to prewriting, became more positive in their reviews, and came to see composing processes as recursive, often unpredictable, and always experimental. Each claim was accompanied by hand-drawn pre-and-post survey figures, complete with *F*-tests levels of statistical significance penciled in. Toward the end of the report, Nystrand offered a theoretical basis for the value of peer review. "In Vygotskian terms," he wrote, "we may regard intensive peer review as a

formative social arrangement in which writers become consciously aware of the functional significance of composing behaviors, discourse strategies, and elements of text by managing them all in anticipation of continuous reader feedback" (p. 12). As to the usefulness of his research for others, Nystrand was pedagogically forceful. Requiring careful classroom planning, peer review works best when fully integrated into the curriculum. To establish effective peer review, he noted that the instructor must help students understand what types of group interaction will help them learn to write—and what will not.

With three decades between us, we realize the significance Nystrand's final report holds for researchers in 2018. Wise use of external funding, multimethod research design, claims supported by evidence, and attention to generalizable findings—these are the foundations of the program of research we present in this article. Extending the foundational research of Nystrand, the present study benefits by advancements made in the time since Nystrand used a ruler to draw his figures: articulation of the writing construct; reconceptualization of foundational measurement categories; and technological advances and demands vis-à-vis course design.

Our goal in this article is to broaden the discussion around students and the peer review process, acknowledging the different strands of research that inform this study. There are a number of related sources of evidence that should be examined in order to understand the results we obtained. Rather than narrowing our scope of inquiry, our strategy is to expand the theoretical lens through which we view the writing process.

In this article, we begin with a review of these new developments in writing studies as they are enacted in peer review research with a sample of 837 students enrolled in first-year composition at a public research university. We then turn to a demonstration of the value of construct articulation, the usefulness of fairness as an integrative measurement framework, and the advantages of research in digital environments. Based on findings from our National Science Foundation funded research, we conclude with recommendations for future peer review research.

2.0 Literature Review

A great deal changed in United States education following the 1984 submission of Nystrand's report. The National Institute of Education was abolished the following year and would, in 2002, eventually become the present Institute of Education Sciences. First published in 1981, the cognitive process theory of writing advanced by Linda Flower and John R. Hayes would take hold and evolve into today's widely held social cognitive theory. And while Apple had prophetically announced the Macintosh as the next big thing in a commercial directed by Ridley Scott, the internet as we know it, managed by the first web browser, was to come a decade after Nystrand slipped his final report into the mailbox.

To contextualize the present study, we begin with our articulation of the writing construct. We then turn to our framework for foundational measurement categories and the way these categories are framed within *MyReviewers* (*MyR*), a suite of cloud-based resources and tools designed to leverage writing collaboration. While some of what we present is an extension of

Nystrand's work, other aspects—not necessarily the technological ones—were unimaginable in 1984.

Peer review is widely used in first-year composition programs. Despite its prevalence, writing studies research surrounding the practice of peer review has been scarce (Haswell, 2005, p. 211). Outside of writing studies, in particular educational measurement studies, the research is more prevalent. Studies show that in relation to general writing assignments, peer reviews lead to generally positive outcomes (Moxley & Eubanks, 2016; Ross, Liberman, Ngo, & LeGrand, 2016). After reviewing 109 publications that analyzed peer review, Topping (1998) concluded that peer assessment can lead to an improvement in student grades just as effectively, if not more so, than the teacher assessment. In fact, peer review can influence students positively as they revise, leading to more thorough revisions (Raymond, 1989; Lawrence & Sommers, 1996).

However, the benefit is not one sided; peer reviews do not just help the person being reviewed, but have been shown to impact positively the reviewer and reviewee (Dochy, Segers, & Sluijsmans, 1999). In part, this often-unanticipated consequence is related to metacognition—the ways that peer-review encourages self-reflection on the part of the reviewer (the individual reflecting on the review just given) and reviewee (the individual who has received that review) (Dochy, Segers, & Sluijsmans, 1999). Recent studies have shown the importance of reflection for learning, encouraging the learner to carefully reflect on gained knowledge (Gibson, Kitto, & Bruza, 2016), as well as the importance of writers' ability to discuss their writing strategies or to control a meta-language for writing (Jarratt et al., 2009; Meizlish et al., 2013). Peer review can help reviewers identify misconceptions they may have about the assignment, specifically when they are reviewing a number of student papers. In seeing certain textual elements in others' papers, this recognition can lead the reviewer to consider opportunities for revision in her own work.

Much, if not all, of this research has proceeded in the absence of a defined, well-articulated view of the writing construct. While it is true that the social cognitive theory of writing remained in force during all of the research noted above (Flower & Hayes, 1981), it is equally true that construct conceptualization defined revision as a task undertaken in isolation by the writer. In 2012, Hayes defined revision as follows: Revising written text is "best thought of as a specialized writing activity. Revising is typically initiated by the detection of a problem in an existing text. It involves planning a solution to the problem (in written form or not), translating that solution into language, and transcribing that language into new text to replace the old text" (p. 376). It is not that this definition is incorrect but, rather, that it is limited to a linear and machine model of task construction, problem identification, and problem resolution conducted without the benefit of others. Expanding that model in 2014, Leijten, Van Waes, Schriver, and Hayes reported on the construction of documents using digital sources and identified activities—digital searching, visual content, and managed attention and motivation over multiple tasks—not found in the present writing models. The need to collaborate was identified as central to professional writing.

Captured, therefore, was a need to expand models of the writing construct based on cognitive, interpersonal, and intrapersonal demands. Simply put, in this model, peer review is not only a pedagogy enacted to improve writing; rather, the metacognitive nature of peer-review is part of the construct itself. Such is the view advanced in this article.

The origin of this radical extension of the writing construct is less than five years old. Influenced by the National Research Council (2012), White, Elliot, and Peckham (2015, Figure 3.1, p. 75) have offered a three domain model that is the basis of the research described in this study. Specifically, self-reflection accompanying peer review is understood as a facet of the intrapersonal domain; the review that has been given—a transactional act—is understood as a facet of the interpersonal domain. In the present study, we have used this three domain framework and defined peer review as the process of students reading and evaluating each others' work, an act that includes articulating the criteria used to perform their evaluation. In the present study, we concentrate exclusively on formative review of intermediate drafts—a review that is rubric driven. We seek evidence to support inferences regarding self-reflection (how students evaluate the review they gave) and transaction (how students evaluate the review given by other students).

Investigating fairness in peer-review includes investigating whether any given demographic group disproportionately benefits or loses from the practice. Research on the relationships between peer review and sub-group profiles such as gender and English language learning (ELL) is therefore of critical importance in our understanding of the advancement of student learning. Unfortunately, very few studies examine peer review in relation to gender. Tucker (2014) observed that women received significantly higher marks from their peers, with a statistical difference in peer ratings received by males compared to those received by females ($F(1, 3784) = 15.568$, $p = 0.001$). With a total of 1,523 student participants and 18,814 assessments, Tucker's study is one of the largest published analyses addressing gender. In a study of 182 student peer reviewers, Hamer et al. (2015) found that there were only stylistic differences between male and female peer reviewers: Women gave more general comments and more frequently wrote in a personal voice (p. 161). These studies also rarely, if ever, take into account those who identify as non-binary, further limiting their results.

In comparison to the sparse research on gender, more research has been conducted on peer review and ELL students. Chang's (2016) review of the last two decades of second language (L2) peer review research notes that ELL students welcomed peer review as an addition to, rather than instead of, instructor feedback (p.86). Lundstrom and Baker (2009) found that L2 students providing peer reviews had a greater improvement in their drafts than did those who received peer reviews. Hu and Lam (2010) investigated the effectiveness of peer review with English L2 learners from China, finding that improvements to the revision were linked to peer reviews. Similarly, Paulus (1999) found that students' revisions based on instructor and peer feedback were more thorough than the changes they made on their own. Liang's (2010) study incorporated online peer interaction, but focused more on group discourse, rather than peer review of one another's papers. Recently, Leijen (2017) has examined the use of online peer review for second

language students, with attention to the types and traits of feedback and how these influence revisions made in subsequent drafts.

As the literature illustrates, studies examining peer review through sub-group analysis (gender, ethnicity, race, and English language learning) are very recent indeed. In order to conduct such studies in a principled framework, evidence is best gathered under the foundational measurement categories of fairness (the validity of score interpretation and use for individuals and sub-groups; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 49), reliability (consistency across instances of the assessment procedure; AERA, APA, & NCME, 2014, p. 33), and validity (the degree to which evidence and theory support interpretation of scores for their proposed uses; AERA, APA, & NCME, 2014, p. 11).

We include fairness in our testing to better align ourselves with current research. The 2014 publication of the *Standards for Educational and Psychological Testing* opened the door, elevating fairness to be its own standard, of equal value to validity and reliability, to be included in educational research. In order to complete a robust study, we must ask ourselves how students feel about their reviews, in an attempt to ensure that all students have equal opportunity in the writing classroom. As recent work demonstrates, fairness is no longer peripheral, but is now central, an integrative force, in the study of reliability and validity (Elliot, 2016; Kelly-Riley & Whithaus, 2016; Poe & Inoue, 2016). Under this framework, fairness is defined as the identification of opportunity structures created through maximum construct representation under conditions of constraint—and the toleration of constraint only to the extent to which benefits are realized for the least advantaged. We have used this integrative framework in the present study.

Online peer review is a recently new addition to writing programs. Among the systems designed by classroom instructors to support revision through peer review are the following: ELI (Hart-Davidson, McLeod, Klerkx, & Wojcik, 2010), *MyR* (Moxley & Eubanks, 2016), SWoRD™ (Falakmasir, Ashley, Schunn, & Litman, 2014), and WriteLab (Hewett, 2015). To varying degrees, each of these systems leverages a wide variety of learning analytic techniques such as the ones reported in this article to structure opportunities for student learning.

Within these environments, the study of online peer review has often been contested as to its classroom value (Wen & Tsai, 2006; Willey & Gardner, 2010). Recently Wilson, Diao, and Huang (2015) investigated student opinion regarding online peer review. They found students disliked the extra amount of work online peer assessment required, responding with "frustration with the perceived unfairness of peer assessment review" (Wilson, Diao, & Huang, 2015, p. 28). Similarly, Wen and Tsai (2006) found that while students generally held positive attitudes towards peer review, they were less positive about online peer review, in part because students felt the online aspect was a mere technical tool rather than a method to advance writing improvement. There have been few studies that investigate students' perceptions of peer review generally (Cheng & Warren, 1997; McLaughlin & Simpson, 2004; Struyven et al., 2006), but none of these have been on a large scale. Of the few studies on peer review in digital environments, recent research by Moxley and Eubanks (2016) examines the relationship between

peer and teacher assessments using 46,689 reviews, finding low-to-modest correlations between peer ratings and instructor ratings on individual assignments. In order to complete such an extensive study, Moxley and Eubanks utilized *MyR* to collect a large number of samples. Use of *MyR* is central to our inquiry into students' responses to peer review.

As demonstrated in this literature review, there is a broad range of what we do know about peer review, and yet, there are still gaps in our knowledge, particularly when we try to bring these diverse pieces together. What are students' perceptions of the peer review process, when students from diverse backgrounds in a first-year composition class complete peer reviews online? While research interrogates each of these elements, few studies, if any, examine these pieces together. These are the gaps that our study attempts to begin to fill, as demonstrated by our research questions.

3.0 Research Questions

We use three categories of questions in the present study. Taken together, the following questions are designed to produce information on a specific domain of the writing construct, on evidence-based claims, and on generalization of our findings to other instructional sites.

3.1 Construct Modeling: Intrapersonal and Interpersonal Domain

1. In the study of self-reflection (i.e., the value students place on their own reviews), we ask two questions of this facet of the intrapersonal domain:

    A. How may we describe the quality of student self-reflection?

    B. How may we describe the tone of student self-reflection?
2. In the study of transaction (i.e., the value students place on reviews received from other students), we ask two questions of this facet of the interpersonal domain:

    A. How may we describe the quality of the feedback the student has been given?

    B. How may we describe the tone of the feedback the student has been given?

3.2 Measurement Question: Forms of Evidence

3. In the study of evidence related to fairness, we ask the following: What may be learned by investigation of responses related to self-investigation and transaction when student sub-groups are disaggregated according to gender, ethnicity, race, and English language learning?
4. In the study of evidence related to reliability, we ask the following: What forms of evidence related to response consistency are useful in the analysis of peer review in terms of responses related to self-reflection and transaction?
5. In the study of evidence related to validity, we ask the following: How does a precise definition of the writing construct lend precision to construct-related evidence of self-reflection and transaction in peer review?

3.3 Digital Learning Question: Affordances

6. In the study of digital learning, we ask the following: What is the instrumental value of Research Questions 1 to 5 in terms of affordances—of the *MyR* peer review process?

3.4 Research Design

As part of a program of research supported by the National Science Foundation, our research design benefits from architectures of principled investigation embedded in two related frameworks. The first, Evidence-Centered Design (ECD), is a powerful research framework designed to help multidisciplinary teams of experts develop common language, mental models, design artifacts, and best practices (Mislevy, Steinberg, & Almond, 2003). The second, Design for Assessment (DFA), is a dynamic conceptual model that allows postsecondary institutions to identify the variables that impact a writing program and to ecologically model the variables to increase student success. The framework advances a component design emphasizing consequence, theorization, standpoint, research, documentation, accountability, sustainability, processes, and communication (White, Elliot, & Peckham, 2015).

To lend specification to these frameworks, the design of the present study is presented in terms of sampling plan design, use of digital platform, process of peer review under investigation, the curriculum in which the peer review practices are embedded, elements of the survey, targets of evidence, form of statistical analyses, data retrieval and storage, and adherence to Institutional Review Board (IRB) procedures. While this section of the article may seem unnecessarily elaborate, a detailed discussion of design is important to understanding the value of the present study as but one instance of what the future holds for digital learning.

3.5 Sampling Plan

In this study, we examine students' ratings of the peer reviews they have given themselves (self-reflection) in relation to the ratings their peers have given (transaction). We focus in this study solely on the answers given to the questions shown in Figures 2 and 3 as they were distributed in ENC 1101 in the fall 2016 semester at the University of South Florida at Tampa (USF). The study examines the responses of 837 students. Of these, there is information from 832 students who provided information on self-reflection and 835 students who provided information on transaction. A description of these students is provided in Table 1.

Table 1

Sampling Plan (n = 837)

Gender	Ethnicity	Race	First language	Required admission and placement scores	Course interim project rubric scores
Male = 243	Non-Hispanic = 362	American Indian or Alaska Native = 3	English = 340	ACT English = 17 Reading = 18	Project 1 *M* = 2.94 *SD* = .72 Range = 1, 4
Female = 265	Hispanic = 110	Asian = 65	English & other = 38	SAT Evidence-Based Reading and Writing = 440	Project 2 *M* = 3.01 *SD* = .7 Range = 1, 4
Transgender = 1	Do not wish to answer = 49	Black or African American = 52	Other = 130	TOEFL iBT® total score = 79	Project 3 *M* = 3.01 *SD* = .66 Range = 1, 4
Other = 2	Not answered = 316	Native Hawaiian or other Pacific Islander = 1	Do not wish to answer = 13		
Do not wish to answer = 19		White = 284	Not answered = 316		
Not answered = 316		More than one race = 32	Total = 832		
		Do not wish to answer = 81			
		Not answered = 316			

As Table 1 shows, the student population of the course is diverse, generally reflecting the 41% non-white students. Admission to the course requires mid-range standardized test scores. While not a part of the present study, rubric performance scores on the three course projects discussed below, awarded on a 4-point range, with 4 as the highest score, reveal that students perform well on the three course projects described below.

This student total allows generalization inferences to be made to the all students in ENC 1101 for that fall 2016 semester. Further, if an 80% confidence level is used with a confidence interval of 2.7, then the 837 students exceed the total of 562 students required to make generalization inferences regarding the 2,465 students admitted in the fall 2016 to the USF Tampa campus in terms of the intrapersonal and interpersonal domains studied. We will return to the concept of generalization inferences in the discussion section of the paper.

3.6 Digital Platform

MyR is a suite of cloud-based resources and tools designed to leverage collaboration. Specifically, these tools aim to improve students' writing, critical thinking, and collaborative competencies by helping instructors and students provide more useful, explicit feedback on student writing. *MyR* enables text feedback using PDF markup tools, including more than 200 Community Comments (a library of comments that serve as a multimedia English handbook, with articles, videos, and *try it* exercises available for most comments). When peer reviews are enabled, students and instructors are able to view the collective written feedback (in-text sticky notes, rubric box comments, and Community Comments) given to a paper on one page. Thus, the student is able to analyze the feedback in an accessible way in order to create a revision plan. The software is currently used at several universities across the United States; the University of South Florida has been employing *MyR* since the spring semester of 2009.

Using the product of *MyR*, students upload their papers, and students and instructors provide feedback electronically. In the fall 2016 semester, USF offered three delivery forms of ENC 1101: a traditional face-to-face class, a completely online class, and a "workshop model," in which students meet in the traditional class period one hour a week and then in student-led peer review groups on the second day for about twenty minutes. Classes are capped at 22 students.

3.7 Peer Review Process

In the *MyR* platform, peer reviews are assigned in groups. Typically group size varies from two to five students. If group size is two students, there will be only one record in the data. If group size is five, there will be four records in the data. Each student reviews the submissions of other students in the group. The group size and group members will vary from one project to other. To understand the potential for large response sets, let us take 22 as the average class size. For self-reflection, the calculation is straightforward, with one reflective review per student. At the end of the three intermediate projects, a total of 66 reviews would be available for the class. However, for transaction, the case is much different in terms of probability. If we take three as the normal number of times each project is reviewed by those 22 students, then there are 2,024 possible

combinations for each project. At the end of the semester then, there are 6,072 possible combinations of reviewers and papers.

Because of such expansive variation, multiple records for same peer reviewer yields more than double the number of students. So, for example, answers to the self-reflection helpfulness question (n = 832 students) provide 4,803 responses, shown in Figure 4. Answers to the transaction helpfulness question (n = 835 students) provide 4,809 responses, shown in Figure 5.

3.8 Curriculum

Three projects are required of all sections of ENC 1101. Project 1 asks students to write an annotated bibliography on a single topic or historical figure that will consist of six entries of 200 words each. Three sources must be published between 2000-2010, and three between 2011-2016. Using a total of four sources, Project 2 asks students to write an 800-1,000 word academic essay that argues the ways in which a scholarly conversation about a chosen topic or historical figure has changed—or not changed—over a period of time. Project 3 asks students to create an arguable claim and write an academic essay to support that claim. This essay provides background on the topic or historical figure (context), an arguable claim, evidence to support the claim, counterarguments, and a conclusion that offers the reader directions for further thought. After writing the academic essay, students produce a Google Slide presentation (regarded as a further digital medium) that retains the same purpose and claim as the traditional academic essay.

Students are required to write three drafts for each assignment: 1) an initial draft, usually in the form of an outline, which the student and instructor typically discuss in one-on-one conferences; 2) an intermediate draft that the instructor and students review individually; and 3) a final draft that only the instructor reviews. Peer reviews occur on the intermediate drafts of each assignment; these take place anonymously online using *MyR*.

The students are required to provide five specific comments, corresponding with a highlight they create within the text; two comments using a comment bank provided in the software; and an end comment addressing the paper as a whole. Community Comments are also used during this stage of the review. Students are encouraged by their instructor (and the rubric reinforces this) to address global issues within the paper, such as crafting a strong argument, using peer-reviewed sources effectively, and having effective essay organization. These peer reviews are graded by the instructors, using the rubric shown in Figure 1. On one page, instructors are able to view all of the written comments and Community Comments given by the student; a link is available for the instructor to view the paper highlighted with the comments. At the bottom of the aggregated screen is an end comment box where the instructor is encouraged to respond to the quality, quantity, and tone of the review.

Criteria	Emerging	Developing	Mastering
Quantity	• Missing or fewer than 5 in-text comments • Missing or fewer than 3 Rubric Criteria comments • Missing or fewer than 2 Community Comments • Missing or incomplete Overall Comment	• Fewer than 5 in-text comments • Fewer than 3 Rubric Criteria comments • Fewer than 2 Community Comments • Fewer than 75 words in Overall Comment	• At least 5 in-text comments • At least 3 Rubric Criteria comments • At least 2 Community Comments • At least 75 words in Overall Comment
Quality	• Comments minimally useful, specific, and/or relevant • Comments minimally accurate and not textually grounded • Comments minimally provide few or no constructive suggestions for revision • Community Comments minimally relevant to problems with text • Community Comments minimally address patterns of error • Overall Comment minimally summarizes patterns of error • Overall Comment minimally offers constructive suggestions for improvement	• Comments somewhat useful, specific, and/or relevant • Comments partially accurate and somewhat textually grounded • Comments inconsistently provide constructive suggestions for revision • Community Comments somewhat relevant to problems with text • Community Comments partially address patterns of error • Overall Comment partially summarizes patterns of error • Overall Comment partially offers constructive suggestions for improvement	• Comments useful, specific, and relevant • Comments accurate and textually grounded • Comments consistently provide constructive suggestions for revision • Community Comments relevant to problems with text • Community Comments adequately address patterns of error • Overall Comment adequately summarizes patterns of error • Overall Comment adequately offers constructive suggestions for improvement
Tone	• Comments actively inappropriate, unsupportive, disrespectful, and/or offensive • Comments reflect an informal tone unsuitable for academic writing	• Comments not actively inappropriate, but only partially demonstrate support, respect, and/or empathy • Comments reflect a somewhat casual tone generally not suited for academic writing	• Comments actively appropriate, supportive, respectful, and empathetic • Comments reflect an appropriately formal tone suitable for academic writing

Figure 1. Peer-review rubric.

The peer reviews' usefulness for student writing is reinforced by the revision plan that each student is to write. Utilizing each of their peers' comments, along with the instructors', the students are required to write a three-part revision response, with the first part summarizing the feedback, the second part analyzing what of the feedback was useful and what the student will ignore, and the final part outlining the steps the student will take to revise their assignment for the final draft. Just as the peer reviews are graded, so too are the revision plans. It is therefore important to realize that the process of peer review is, in effect, a constructed response (Bennett, 1993)—a task that asks students to respond to a given set of requirements in order to demonstrate specific abilities that are, in turn directly related to the writing construct.

3.9 Survey Design

Two survey opportunities, given on each of the three projects, have been incorporated within curriculum described above and delivered by the *MyR* platform. As Figure 2 shows, when the student completes a peer review, the first survey appears, asking the student to reflect on the quality of the peer review they have given. Questions center on quality and tone, key elements for peer review (Hamer et al., 2015). In Question 1, quality is labeled as helpfulness, and the words used to describe each number on the 1-5 scale reflect language in the rubric. Students are

then asked to rate their tone in terms of politeness (Question 2), kindness (Question 3), and encouragement (Question 4).

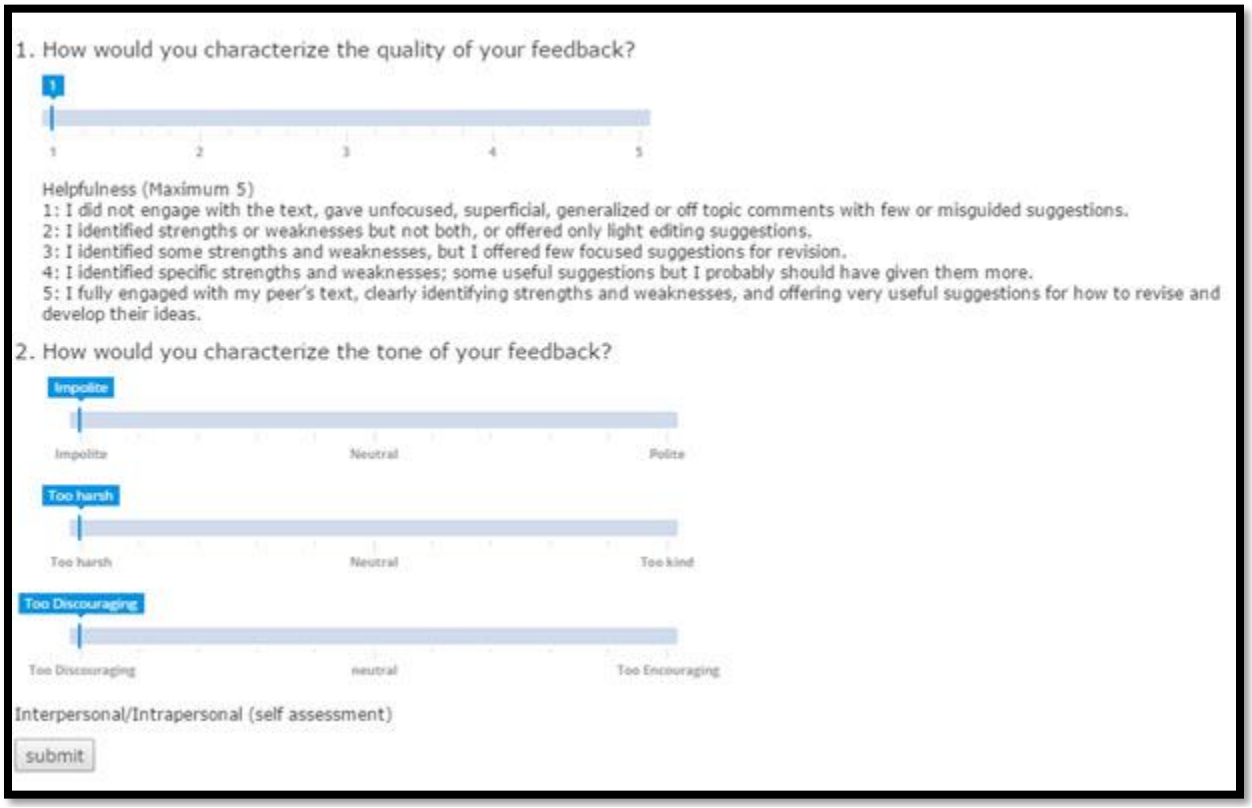

Figure 2. Self-reflection survey question.

The second survey, shown in Figure 3, is available to students when they complete the revision plan. The revision plan is shown on an interface where students can see the individual peer reviews and the feedback from the instructor on one screen. There are then links to connect to the paper format version, complete with in-text highlights and comments, of each review. For each peer review on the three projects, a survey is offered, giving students the opportunity to evaluate the transactional nature of the feedback they have received. The questions mirror those asked of the student after reflectively completing their peer reviews, and the order of the questions is the same on the transaction survey.

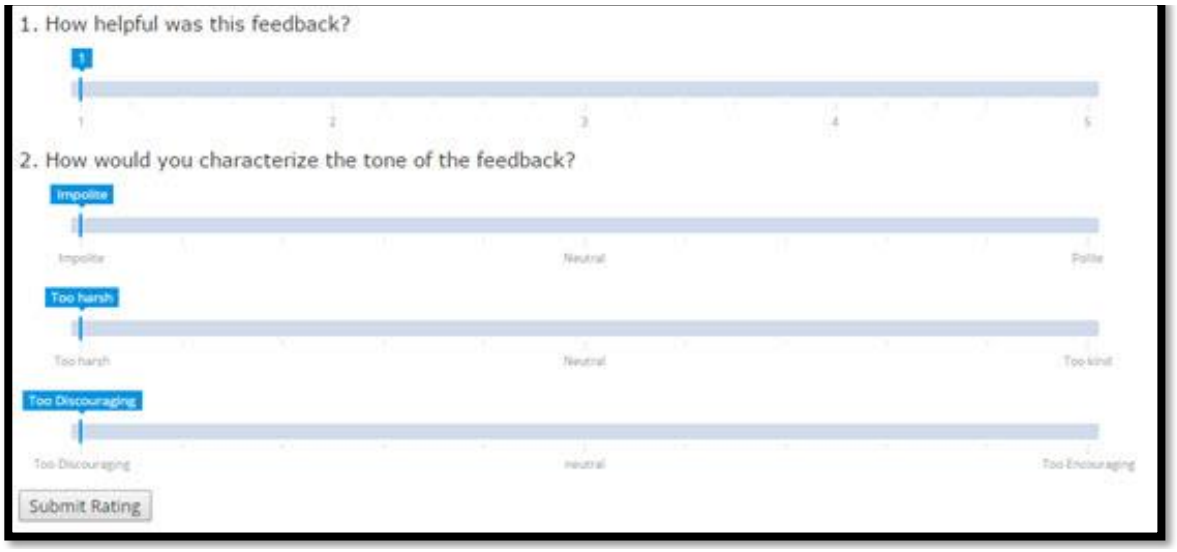

Figure 3. Transaction survey question.

3.10 Evidential Analysis

We have used three evidential targets in our search for information contained in the two surveys: fairness (as the integrative principle), validity, and reliability.

3.10.1 Fairness. In the present study, fairness evidence is examined through survey response disaggregation by gender, ethnicity, race, and English language learning. Evidence of fairness is shown in Tables 2 and 3. As an occasion of opportunity to learn (Moss, Pullin, Gee, Haertel, & Young, 2008) about the students in the study and to create new opportunity structures for student success based on intrapersonal and interpersonal domain knowledge, response disaggregation by sub-group yields both group and individual student investigation.

3.10.2 Validity. As a category of evidence, construct representation has received special attention in writing assessment, with locally-developed assessments perceived as yielding robust presence of the writing construct (Behizadeh & Engelhard, 2015; Condon, 2013). In the present study, evidence regarding the internal structure of the surveys is understood as evidence of model strength of both self-reflection and transaction associated with, respectively, intrapersonal and interpersonal elements of the writing construct. Following the *Standards for Educational and Psychological Testing*, (AREA, APA, & NCME, 2014), we understand that the rationale for our interpretation of the surveys rests on the relationships among the responses. This study reports on internal structure evidence in the form of correlation coefficients and probabilistic analysis. Evidence of internal structure of the surveys is shown in Tables 4 and 5.

3.10.3 Reliability. Consistency for the two surveys is established by response consistency in both surveys across the three course projects. This study reports on reliability evidence in the form of correlation coefficients and probabilistic analysis. Evidence of survey reliability is shown in Tables 6 and 7.

3.11 Statistical Analyses

Basic descriptive and inferential analyses are used throughout the study, with inferential statistics employing probabilistic models. For ANOVA analyses shown in Tables 3 and 4 in terms of analysis of race differences, the Bonferroni correction was used. The correlation ranges used in analyses and discussions are as follows: high positive correlations = 1.0 to 0.70; medium positive correlations = 0.69 to 0.30; and low positive correlations = 0.29. Data was drawn from *MyR* in Excel 2016 and analyzed in SPSS 22. However, as noted below, statistical packages such as SPSS are limited in dealing with massive (i.e., big) data analysis (National Research Council, 2013).

3.12 Data Retrieval and Storage

Technically, *MyR* operational data is stored in a Microsoft SQL Server database, which immediately reflects all live changes. Data is split among multiple tables as required by the database normal forms to avoid redundancy and anomalies and to increase integrity. Such data organization greatly facilitates data insertions, deletions, and updates frequently occurring in live databases.

Such data organization, however, does not work well for data retrievals required to extract datasets for analytical studies. Following a common approach, we designed a data warehouse and correspondingly de-normalized and joined multiple tables into a few larger ones, which were subsequently stored on a dedicated research server in another database. We thus used Microsoft SQL Server 2016 Developer Edition that allowed free usage for non-production systems. Using a free version of Microsoft Visual Studio Community Edition and C# language, we then designed a custom application to extract, transform, and load the data into the warehouse. Furthermore, this application was extended to allow de-identification as required by IRB and extracting datasets according to multiple parameters, such as university, semester, and course major.

Our work revealed significant slowdowns and obstacles appearing while processing big data in a relational database. Known as the four Vs—volume, variety, velocity and veracity—of big data, these demand categories require that the processing of big data must be accompanied by special tools not only for its large size but also because of multiple formats and the uncertainty of human error in processing data quality. Traditional data processing tools—spreadsheet software, relational databases, or statistical packages—are not capable of effective processing of this scale.

Massive parallel computations employing clusters of computers are therefore needed to effectively overcome this problem. An example of such an ecosystem is Apache Hadoop, allowing software developers and data scientists to exercise data analytics at a large scale (White, 2015). To maintain state-of-the-art data retrieval and storage, our final transition will be to MongoDB, a document-oriented, scalable data store used by 30 out of 100 of the world's largest organizations (Mongo, 2016).

3.13 IRB Procedures

The survey has been approved under the IRB of the University of South Florida and other institutions affiliated with this study. Surveys are optional for students to complete, and the data only comes from those students who have opted in for the research.

4.0 Results

Results are presented in terms of the research questions. Research Questions 1 and 2, with analyses of student responses, are dealt with comparatively. Because of information complexity, results from Questions 3, 4, and 5 are analyzed individually by construct. Question 6 returns to the comparative analysis. When response differences are recorded, they meet or exceed the .05 level of statistical significance. When there is no difference, exact *p*-values are recorded in the referenced table.

4.1 Overall Descriptions: Self-Reflection and Transaction

The research incorporates survey responses from 837 students, totaling 4,803 self-reflection surveys (from 832 students on three assignments) and 4,309 from transaction surveys (from 835 students on three assignments). We have provided the response distribution for self-reflection and transaction in Figures 4 and 5. To provide additional detail, Table 2 includes the questions and sub-group analysis for the self-reflection survey. Table 3 provides similar additional information on the transaction survey.

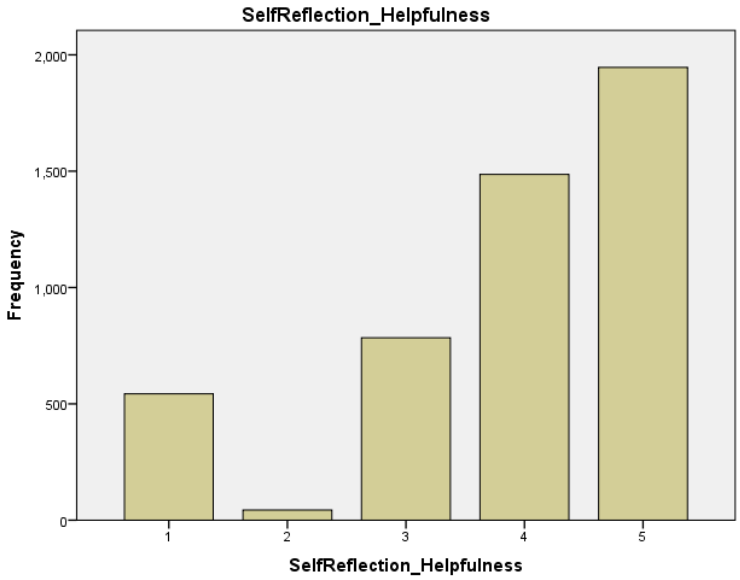

Figure 4. Self-reflection score distribution: Helpfulness (n = 832 students providing 4,803 responses).

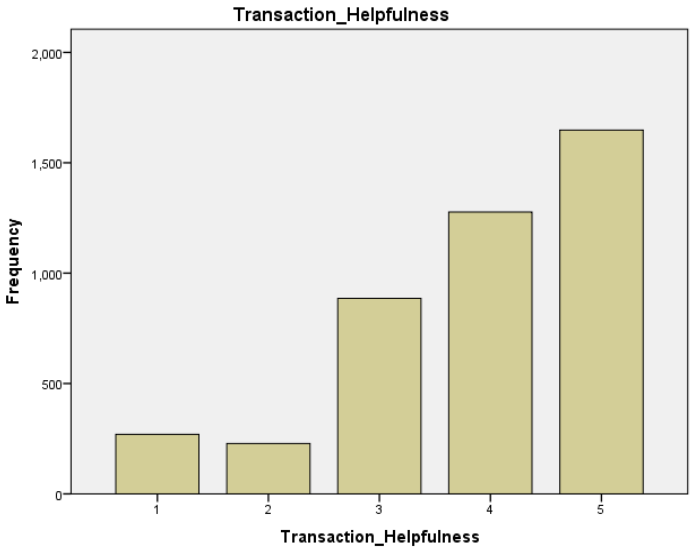

Figure 5. Transaction score distribution: Helpfulness (n = 835 students providing 4,809 responses).

Table 2

Self-Reflection Survey (n = 832 students providing 4,803 responses)

Questions	Total number of responses and missing responses	*M* *SD* Range	Male n *M* *SD* Range	Female n *M* *SD* Range	Gender difference (*t*, *df*, *p*)	Non-Hispanic n *M* *SD* Range	Hispanic n *M* *SD* Range	Ethnicity difference (*t*, *df*, *p*)	White n *M* *SD* Range	Asian n *M* *SD* Range	Black n *M* *SD* Range	Race difference (*F*, *df*, *p*)	English as only language n *M* *SD* Range	English and another language n *M* *SD* Range	English language learning difference (*t*, *df*, *p*)
1. How would you characterize the quality of your feedback in terms of helpfulness?	4,803 235	3.88 1.27 (1, 5)	1,332 3.83 1.25 (1, 5)	1,567 4.00 1.17 (1, 5)	*t*(2897) = 3.75 *p* < .001	2,107 3.91 1.24 (1, 5)	631 4.07 1.06 (1, 5)	*t*(2736) = 2.91 *p* < .001	1,623 3.99 1.22 (1, 5)	413 3.93 1.14 (1, 5)	316 3.97 1.08 (1, 5)	*F*(2, 2349) = .40 *p* = .67	1,957 3.95 1.22 (1, 5)	223 4.13 .87 (1, 5)	*t*(2178) = 2.19 *p* < .05
2. How would you characterize the tone of your feedback in terms of politeness?	4,803 235	2.25 .64 (1, 3)	1,332 2.26 .64 (1, 3)	1,567 2.33 .617 (1, 3)	*t*(2897) = 3.21 *p* < .001	2,107 2.31 .56 (1, 3)	631 2.32 .64 (1, 3)	*t*(2736) = .52 *p* = .61	1,623 2.32 .634 (1, 3)	413 2.23 .56 (1, 3)	316 2.42 .58 (1, 3)	*F*(2, 2349) = 8.56 *p* < .001 A<B *p* < .001 W<B *p* < .01 A<W *p* < .01	1,957 2.30 .63 (1, 3)	223 2.28 .51 (1, 3)	*t*(2178) = .681 *p* = .68

Questions	Total number of responses and missing responses	*M* *SD* Range	Male n *M* *SD* Range	Female n *M* *SD* Range	Gender difference (*t*, *df*, *p*)	Non-Hispanic n *M* *SD* Range	Hispanic n *M* *SD* Range	Ethnicity difference (*t*, *df*, *p*)	White n *M* *SD* Range	Asian n *M* *SD* Range	Black n *M* *SD* Range	Race difference (*F*, *df*, *p*)	English as only language n *M* *SD* Range	English and another language n *M* *SD* Range	English language learning difference (*t*, *df*, *p*)
3. How would you characterize the tone of your feedback in terms of kindness?	4,803 235	1.94 .44 (1, 3)	1,332 1.96 .46 (1, 3)	1,567 1.97 .412 (1, 3)	*t*(2897) = .792 *p* =.43	2,107 1.96 .37 (1, 3)	631 1.99 .45 (1, 3)	*t*(2736) = 1.42 *p* = .16	1,623 1.96 .421 (1, 3)	413 2.01 .47 (1, 3)	316 2.01 .34 (1, 3)	*F*(2, 2349) = 4.18 *p* < .01 W<B *p* < .05	1,957 1.95 .42 (1, 3)	223 2.0 .33 (1, 3)	*t*(2178) = .16 *p* = .11
4. How would you characterize the tone of your feedback in terms of encouragement?	915 4,123	1.74 .97 (1, 3)	267 1.86 .99 (1, 3)	255 1.91 .10 (1, 3)	*t*(520) = .555 *p* =.52	400 1.87 .99 (1, 3)	78 2.03 .99 (1, 3)	*t*(476) = 1.26 *p* = .21	271 1.85 .99 (1, 3)	97 2.24 .98 (1, 3)	32 1.75 .98 (1, 3)	*F*(2, 397) = 6.15 *p* < .001 B<A *p* < .05 W<A *p* < .001 B<W *p* < .001	310 1.74 .97 (1, 3)	19 2.16 1.02 (1, 3)	*t*(327) = 1.81 *p* = .07

Table 3

Transaction Survey (n = 837 students providing 4,309 responses)

Questions	Total number of responses and missing responses	*M* *SD* Range	Male n *M* *SD* Range	Female n *M* *SD* Range	Gender difference (*t*, *df*, *p*)	Non-Hispanic n *M* *SD* Range	Hispanic n *M* *SD* Range	Ethnicity difference (*t*, *df*, *p*)	White n *M* *SD* Range	Asian n *M* *SD* Range	Black n *M* *SD* Range	Race difference (*F*, *df*, *p*)	English as only language n *M* *SD* Range	English and another language n *M* *SD* Range	English language learning difference (*t*, *df*, *p*)
1. How helpful was the feedback you received?	4,309 729	3.88 1.16 (1, 5)	1, 205 3.91 1.14 (1, 5)	1, 416 3.97 1.15 (1, 5)	*t*(2619) = 1.30 *p* = .20	1, 915 3.95 1.16 (1, 5)	555 3.98 1.09 (1, 5)	*t*(2468) = .43 *p* = .67	1, 475 4.03 1.08 (1, 5)	359 3.78 1.25 (1, 5)	270 3.87 1.18 (1, 5)	*F*(2, 2102) = 8.04 *p* < .001 A<W *p* < .001	1, 776 3.92 1.16 (1, 5)	196 4.02 1.09 (1, 5)	*t*(1960) = 1.12 *p* = 2.7
2. How would you characterize the tone of the feedback you received in terms of politeness?	4,309 729	2.35 .57 (1, 3)	1, 205 2.37 .57 (1, 3)	1, 416 2.33 .56 (1, 3)	*t*(2619) = 1.53 *p* = .13	1, 195 2.35 .57 (1, 5)	555 2.65 .56 (1, 5)	*t*(2468) = .07 *p* = .99	1, 475 2.37 .56 (1, 3)	359 3.37 .56 (1, 3)	270 3.87 .56 (1, 3)	*F*(2, 2102) = 5.82 *p* < .001 A<B *p* < .01 W<A *p* < .001	1, 776 2.33 .57 (1, 3)	196 2.4 .58 (1, 3)	*t*(1960) = 1.66 *p* = .10

Questions	Total number of responses and missing responses	*M* *SD* Range	Male n *M* *SD* Range	Female n *M* *SD* Range	Gender difference (*t*, *df*, *p*)	Non-Hispanic n *M* *SD* Range	Hispanic n *M* *SD* Range	Ethnicity difference (*t*, *df*, *p*)	White n *M* *SD* Range	Asian n *M* *SD* Range	Black n *M* *SD* Range	Race difference (*F*, *df*, *p*)	English as only language n *M* *SD* Range	English and another language n *M* *SD* Range	English language learning difference (*t*, *df*, *p*)
3. How would you characterize the tone of the feedback you received in terms of kindness?	4,309 729	2.04 .37 (1, 3)	1, 205 2.04 .35 (1, 3)	1, 416 2.45 .37 (1, 3)	*t*(2619) = .211 *p* = .83	1, 195 2.04 .37 (1, 5)	555 2.03 .33 (1, 5)	*t*(2468) = .92 *p* = .36	1, 475 2.04 .35 (1, 3)	359 2.03 .37 (1, 3)	270 2.07 .36 (1, 5)	*F*(2, 2102) = 1.4 *p* = .25	1, 776 2.03 .36 (1, 3)	196 2.06 .387 (1, 3)	*t*(1960) = 1.23 *p* = .23
4. How would you characterize the tone of the feedback you received in terms of encouragement?	589 4,449	2.27 .94 (1, 3)	1, 205 2.38 .92 (1, 3)	1, 416 2.16 .99 (1, 3)	*t*(374) = 2.16 *p* < .01	265 2.25 .97 (1, 3)	64 2.19 .99 (1, 3)	*t*(327) = .48 *p* = .63	188 2.34 .94 (1, 3)	52 2.12 1.0 (1, 3)	33 2.45 .91 (1, 3)	*F*(2, 270) = 1.4 *p* = 1.57 *p* = .21	233 2.23 .98 (1, 3)	35 2.31 .96 (1, 3)	*t*(266) = .49 *p* = .62

The self-reflection surveys reveal the majority of reviews (n = 1946) reveal a response of 5 on the helpfulness scale. Shown in Table 2, the mean score for Question 1, 3.88, indicates that students view themselves as above average in their helpfulness. That is, as Figure 4 shows, the students felt they had identified specific strengths and weaknesses, or they fully engaged the text of their classmates. The transactional surveys shown in Figure 5 are so similar that the bar charts appear nearly identical. Shown in Table 3, Question 1 of the transactional surveys reveals the majority of reviews (n = 1648) award themselves a response of 5 on the helpfulness scale. The mean response, 3.88, indicates that students view the reviews given by other students to be above average.

Tone questions of politeness in Question 3 and kindness in Question 4 retain completion rates similar to that of the helpfulness questions. Shown in Table 2, self-reflection surveys reveal that students felt their own politeness (M = 2.25) and kindness (M = 1.94) were neutral. As Table 3 reveals, questions of politeness (M = 2.35) and kindness (M = 2.04) were similarly neutral on the transactional surveys.

However, strikingly different were the declines in survey responses for Question 3 regarding encouragement. In the self-reflection surveys, response rates dropped to 915 with 4,123 missing responses, compared to only 235 missing responses on the other survey questions shown in Table 2. Similarly, the transactional surveys fell to a response rate of 569 students, with 4,449 missing responses for Question 3, compared to only 729 missing responses on the other survey questions shown in Table 3. Similarly notable, no student in either the self-reflection survey or the transactional survey selected the neutral (response 2) response; repeated checks revealed that there was no software malfunction. In the self-reflection survey, the majority of students (n = 547) worried they had been too discouraging and gave themselves a response of 1. Conversely, in the transactional survey, the majority of students (n = 374) believed that the reviews given by others had been encouraging.

Longitudinal studies across projects are also of interest. In focusing only in Question 1 regarding helpfulness on self-reflection and transactions across the three projects, mean responses and high ratings both continue at a fairly steady rate as the semester progresses. As Figure 6 shows, both mean responses and distribution patterns are nearly identical across the three projects in terms of self-reflection. As Figure 7 shows, these patterns are consistent across assignments regarding the transaction responses.

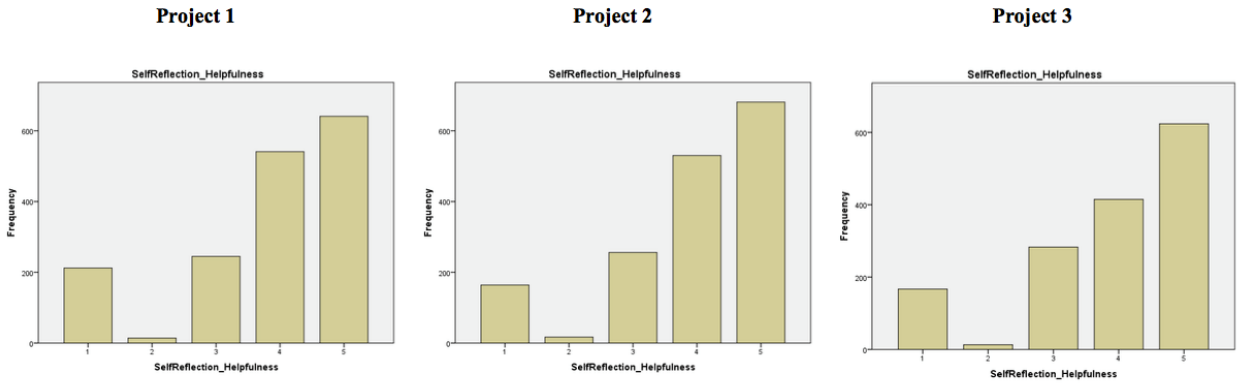

Figure 6. Self-reflection (N = 832 students with 4803 responses): Project 1(n = 1,653, M = 3.84, SD = 1.3), Project 2 (n = 1,648, M = 3.94, SD = 1.23), and Project 3 (n = 1,502, M = 3.88, SD = 1.23)

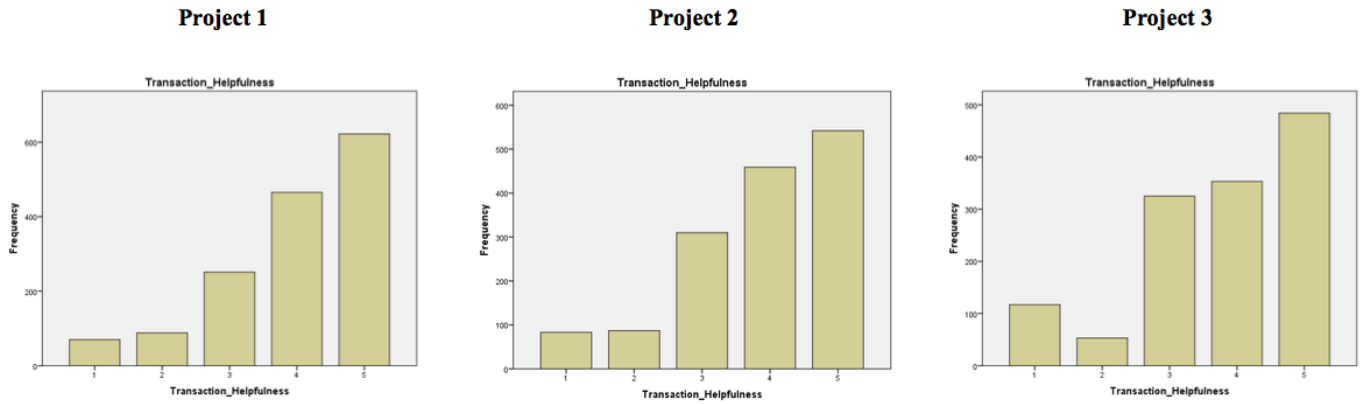

Figure 7. Transaction, Project 1(n = 1,496, M = 3.99, SD = 1.11), Project 2 (n = 1,481, M = 3.87, SD = 1.14), and Project 3 (n = 1,332, M = 3.78, SD = 1.23)

Similarities between Figures 6 and 7 demonstrate no variation in the overall pattern observed in Figures 4 and 5. In terms of participation, self-reflection participants decline at a rate of 9% between Project 1 and Project 3 as the semester progresses as they respond to the helpfulness question. Transaction participants decline at a rate of 11% between Project 1 and Project 3 as the semester progresses.

4.2 Measurement of Self-Reflection: Fairness Evidence

4.2.1 Fairness and gender. Response disaggregation is shown in Table 2 for self-reflection. In terms of gender, women awarded themselves higher responses for helpfulness than men at statistically significant levels. Reading tone, women believed their feedback was more polite than the responses given by men at statistically significant levels. In terms of kindness and encouragement, however, there is no statistically significant difference between men and women.

This survey did not provide an option for students who do not identify as either female or male, thereby limiting our findings.

4.2.2 Fairness and ethnicity. Hispanic students believed their own reviews were more helpful than non-Hispanic students did at statistically significant levels. No statistically significant differences were observed, however, in terms of politeness, kindness, and encouragement of the tone of their own reviews.

4.2.3 Fairness and race. In terms of helpfulness, there is no statistical difference among White, Asian, or Black students. However, statistically significant differences were evident for each group reading politeness of their own reviews, with Black students scoring the highest, followed by White and Asian students. In terms of kindness, Black students believe their reviews are kinder than those of White or Asian students at statistically significant levels. At statistically significant levels, the responses of Asian students reflect their beliefs that their reviews offered more encouragement than the responses of White or Black students, and White students felt their reviews were more encouraging than Black students.

4.2.4 Fairness and English language learning. Those students who claimed proficiency in English and another language felt that their own reviews were more helpful than those students who had English as their only language. There were no significant differences in terms of tone, kindness, and encouragement between the two groups.

4.3 Measurement of Self-Reflection: Reliability Evidence

Table 6 provides reliability estimates of self-reflection between Projects 1 and 2, Projects 2 and 3, and Projects 1 and 3. In each case, for all groups, reliability reaches statistically significant levels. However, for Black students, there is a decline in the level of statistical significance. Levels of correlation are low-to-medium. Notably, for the overall group and for all sub-groups, levels of reliability increase between Projects 1 and 2—and, again, between Projects 2 and 3. Reliability between Projects 2 and 3 are the highest correlations, with all at medium levels. In comparisons of Projects 1 and 3, however, reliability declines in the overall group and in all sub-groups.

Table 6

Self-Reflection: Helpfulness Reliability, All Groups (n = 832 students providing 4,803 responses)

Project	*r*
All	
Project 1 and Project 2	.37**
Project 2 and Project 3	.53**
Project 1 and Project 3	.34**
Male	
Project 1 and Project 2	.40**
Project 2 and Project 3	.52**
Project 1 and Project 3	.34**
Female	
Project 1 and Project 2	.31**
Project 2 and Project 3	.47**
Project 1 and Project 3	.27**
Non-Hispanic	
Project 1 and Project 2	.36**
Project 2 and Project 3	.49**
Project 1 and Project 3	.33**
Hispanic	
Project 1 and Project 2	.33**
Project 2 and Project 3	.51**
Project 1 and Project 3	.21**

Project	*r*
White	
Project 1 and Project 2	.38**
Project 2 and Project 3	.51**
Project 1 and Project 3	.37**
Asian	
Project 1 and Project 2	.25**
Project 2 and Project 3	.57**
Project 1 and Project 3	.35**
Black	
Project 1 and Project 2	.23*
Project 2 and Project 3	.43*
Project 1 and Project 3	.25*
English as only language	
Project 1 and Project 2	.33**
Project 2 and Project 3	.48**
Project 1 and Project 3	.33**
English and another language	
Project 1 and Project 2	.57**
Project 2 and Project 3	.65**
Project 1 and Project 3	.55**

Note. * $p < .05$, ** $p < .01$

4.4 Measurement of Self-Reflection: Validity Evidence of Internal Structure

Strength of the self-reflection model, as determined by the four questions shown in Table 2, is obtained by correlation analysis of the variable with each other. As Table 4 illustrates, helpfulness, politeness, kindness, and encouragement correlate at medium-to-high statistically significant levels. Although not shown, similar medium-to-high correlation rates were found across all sub-groups shown in Table 2, with the exception of low, significant correlations of those students who were English language learners. For those students, statistically significant correlations ($p < .001$) ranged from .22 to .70.

Table 4

Transaction Correlations (Transaction Survey (n = 837 students providing 4,309 responses)

Measures	**1.**	**2.**	**3.**	**4.**	***M***	***SD***
1. Helpfulness (N = 4,809)	—	.45**	.30**	.73**	3.88	1.16
2. Politeness (N = 4,809)		—	.49**	.88**	2.35.	.57
3. Kindness (N = 4,809)			—	.88**	2.04	.37
4. Engagement (N = 4,809)				—	2.27	.94

Note. *$p < .05$, **$p < .01$

4.5 Measurement of Transaction: Fairness Evidence

4.5.1 Fairness and gender. Response disaggregation is shown in Table 3 for transaction. In terms of gender, the reviews given by women were no more helpful than those given by men. The same is true for politeness and kindness. In terms of encouragement, however, men were more encouraging than women in their feedback.

4.5.2 Fairness and ethnicity. Hispanic students provided reviews that were perceived as equally helpful as non-Hispanic students. No statistically significant differences were observed in terms of politeness, kindness, or encouragement.

4.5.3 Fairness and race. In terms of helpfulness, White students were recorded as giving more helpful feedback than Asian students, but there were no differences between Asian and Black reviews. Black students gave more polite reviews than Asian students; in turn, Asian students gave more polite reviews than White students. In both kindness and encouragement, no differences were observed among the four student groups.

4.5.4 Fairness and English language learning. No statistically significant differences were observed in terms of helpfulness, politeness, kindness, or encouragement.

4.6 Measurement of Transaction: Reliability Evidence

Table 7 provides reliability estimates of transaction between Projects 1 and 2, Projects 2 and 3, and Projects 1 and 3. While the overall group and most sub-group responses attain reliability at statistically significant levels, Hispanic and Black Students fail to reach the .05 level in Projects 1 and 3. For ELL students, reliability is achieved only between Projects 1 and 2. Levels of correlation are generally low. For the overall group, males, and students who have English as their only language, levels of reliability increase between Projects 1 and 2—and, again, between Projects 2 and 3. This pattern does not hold across all groups.

Table 7

Transaction: Helpfulness Reliability, All Groups (n = 837 students providing 4,309 responses)

Project	*r*
All	
Project 1 and Project 2	.27**
Project 2 and Project 3	.36**
Project 1 and Project 3	.19**
Male	
Project 1 and Project 2	.12**
Project 2 and Project 3	.29**
Project 1 and Project 3	.16**
Female	
Project 1 and Project 2	.36**
Project 2 and Project 3	.31**
Project 1 and Project 3	.16**

Project	*r*
Non-Hispanic	
Project 1 and Project 2	.31**
Project 2 and Project 3	.24**
Project 1 and Project 3	.18**
Hispanic	
Project 1 and Project 2	.25**
Project 2 and Project 3	.4**
Project 1 and Project 3	.0 nss
White	
Project 1 and Project 2	.29**
Project 2 and Project 3	.31**
Project 1 and Project 3	.11*
Asian	
Project 1 and Project 2	.40**
Project 2 and Project 3	.31**
Project 1 and Project 3	.33**
Black	
Project 1 and Project 2	.31*
Project 2 and Project 3	.31*
Project 1 and Project 3	.08 nss
English as only language	
Project 1 and Project 2	.24**
Project 2 and Project 3	.28**
Project 1 and Project 3	.11*

Project	*r*
English and another language	
Project 1 and Project 2	.27**
Project 2 and Project 3	.15 nss
Project 1 and Project 3	.21 nss

Note. * $p < .05$, ** $p < .0$, *nss* = not statistically significant

4.7 Measurement of Transaction: Validity Evidence of Internal Structure

Strength of the transaction model, as determined by the four questions shown in Table 3, is obtained by correlation analysis of the variable with each other. As Table 8 illustrates, helpfulness, politeness, kindness, and encouragement correlate at medium-to-high statistically significant levels. Although not shown, similar medium-to-high correlation rates were found across all sub-groups shown in Table 3 with the exception of lower ranges of statistically significant ($p < .001$) correlations of Asian, Black, and White students. For Asian students, correlations ranged from .26 to .81. For Black students, correlations ranged from .25 to .94. For White students, correlations ranged from .29 to .93. For students whose sole language was English, correlations ranged from .24 to .90. In each case, high correlations were also retained.

Table 8

Self-Reflection Correlations (n = 832 students providing 4,803 responses)

Measures	1.	2.	3.	4.	*M*	*SD*
1. Helpfulness (N = 4,803)	—	.64**	.61**	.86**	3.88	1.27
2. Politeness (N = 4,803)		—	.63**	.90**	2.25	.64
3. Kindness (N = 4,803)			—	.91**	1.94	.44
4. Engagement (N = 4,803)				—	1.74	.97

Note. *$p < .05$, **$p < .01$

4.8 Digital Learning Question: Affordances

Following Greeno and Gresalfi (2008), we have found that opportunities to learn are best pursued by identifying the affordances for understanding participation and practice within *MyR*. Relational in nature, affordances for an individual within an activity system such as *MyR* include the following: identification of the resources and practices of the system; analysis of access of groups and individuals to those resources and practices; and discovery of dispositions and abilities of group and individual participation in ways that support learning the construct under examination. Tables 2 through 7 provide evidence on fairness, reliability, and validity of the students described in Table 1 to access peer review in its intrapersonal and interpersonal domains.

This is not to say, however, that there are not differences in that access. Notable is the reversal in which many students, operating under stereotype threat (Stricker & Ward, 2004), appear to have broad access to these domains studied. According to Figure 2, answers to the self-reflection questions were translated to numerical values from 1 (no engagement) to 5 (full engagement) for question 1, from 0 (impolite) to 3 (polite) for question 2, from 0 (too harsh) to 3 (too kind) for question 3, and from 0 (too discouraging) to 3 (too encouraging) for question 4. Similarly, numeric codes are shown in Figure 3 for Transaction Survey questions.

For each question in Table 2—unless noted otherwise—total number of respondents *n* is in row 1, mean value *M* is in row 2, standard deviation *SD* is in row 3, and range is in row 4. Table 2 illustrates that the self-reflection of Hispanic students (*M*=4.07) regarding the helpfulness of their reviews is higher than that of non-Hispanic students (*M*=3.91). In similar fashion, Black students (*M*=2.42) feel that their reviews are more polite than White students (*M*=2.32). In terms of transaction, as shown in Table 3, there are no statistically significant differences between the helpfulness (3.95 vs. 3.98), politeness (2.35 vs. 2.65), kindness (2.04 vs. 2.03), or encouragement (2.25 vs. 2.19) of Non-Hispanic and Hispanic students. In terms of politeness, Black students provide more helpful reviews than do their Asian classmates (3.87 vs. 3.78). In many ways, students often found to be low-performing writers thrive in digital learning platforms when the construct itself is expanded to include the metacognitive.

Indeed, one interesting exercise is to compare Tables 2 and 3 to the standardized test scores of the SAT for 2016 college bound seniors in terms of ethnicity, race, and first language learned (College Board, 2016). On the standardized test, we find the same tired comparisons that demonstrate the continued disenfranchisement of many sub-groups. In contrast, the study reported here shows little of those patterns. To those who say that the SAT and *MyR* examine different forms of the writing construct, we agree. Our response poses a different sort of question: Why—with all we are learning about broader views of the writing construct, expanded foundations of measurement, and advantages of digital learning—do we continue assessment practices in ways that merely demonstrate the absence of affordances rather than the presence of student learning?

Answers to Research Questions 1 to 5 may be understood as a demonstration of the instrumental value of the *MyR* peer review process. Far from a narrative of satisfaction, our work

has just begun. It would, for instance, be naive to believe that we have done more than provide little more than a sliver of information about the facets accompanying the interpersonal and intrapersonal domains accompanying peer review with the eight questions examined here. However, based on the response fatigue (percentage of missing responses ranges from 5% to 14%) recorded in Figures 6 and 7 and in Question 4 of Tables 2 and 3, it would be equally naive to suggest that additional questions would be in order. In our search for unobtrusive methods, we have begun examining the comments associated with peer review in order to develop corpora of this little-examined genre. Base-line research by Rudniy and Elliot (2016) reveals that n-gram analytic methods identify important information about instructor and student use of course threshold concepts from rubrics and syllabi as they are used in comments. With special attention to Natural Language Processing and Latent Semantic Analysis methods, this research will lead to analysis of the comment corpora in order to determine effective elements of intrapersonal and interpersonal domains. While research has not yet explored connections between patterned language use in peer review and intrapersonal and interpersonal domains, possibilities include exploring how interpersonal habits are realized in stance patterns that show diplomacy and caution, and how aspects of intrapersonal self-reflections help illuminate what constitutes an effective meta-language for students' talk about their writing (Aull, 2015).

As knowledge of the peer review process expands, *MyR* has the capability of providing real-time actionable analytics. In the same way that the surveys shown in Figures 2 and 3 popped up when students submitted their peer reviews, reports based on the analyses described in this article can appear in real time in order to guide student responses. Think about the students providing the 543 responses shown in Figure 4 who reported beliefs that they had not provided helpful reviews. Now imagine reports that could, at once, provide comparative analysis based on the responses of other students, suggest advice on improving reviews, and identify digital resources that could help with upcoming reviews. These next steps are within grasp.

5.0 Discussion

When Nystrand typed up his final report on the effectiveness of peer review in expository writing instruction in 1984, the ability to draw down and analyze data at the speed noted in the present study was unimaginable. Equally unimaginable was the nearly 470 gigabytes of structured and unstructured data in in textual, numeric, and PDF formats that accompanied the initial 2016 drawdowns of similar data across campuses—an amount of information that no longer allows ordinary spreadsheets and web browsers to function. As the section on data retrieval and storage above suggests, even the software used in the present study is, under certain conditions, useless. And the analysis we have presented here is only a sliver of what is available for further study.

It would, however, be an error to believe the only terrain separating his work and ours is technological in nature. The domain-based models with which we work today were largely unknown. While cognition was to be all the rage in the 1980s, the sociocognitive and sociocultural dimensions of writing—with accompanying domain-based models—were three

decades in the future (Mislevy, 2016). In terms of foundations of measurement, evidentiary reasoning was certainly in force, but the systematic validation accompanying interpretation and use arguments was not (Kane, 2016). And digital environments would, for a long time, be little more than digital filing systems leading to course management systems. Only in the last five years have we witnessed the potential of ELI, MyR, SWoRD™, and WriteLab to allow the sort of analysis we have provided above.

6.0 Conclusions

Now that this research exists, what can we reasonably expect to be done? As Carol Weiss (1998) notes in the case of program assessment, the instrumental use of an evaluation is not always what the evaluators intended. Often, much is dependent on the amount of work required in ordered to implement the findings. Additionally, most writing programs are bound by "rigid limits," whether because of law, regulations, or the mere habit of how the program has always functioned (Weiss, 1998, p.28). How to overcome these barriers?

As a national leader in evaluation and the Beatrice B. Whiting Professor in the Graduate School of Education at Harvard University, Weiss had become dissatisfied with the writing of reports and the publication of research as the final step in assessment. In response, she developed an alternative mode of evaluation, known as theory-based evaluation, in which the evaluation itself (the findings) is based on theories of change (paths for action based on the findings). Put directly, then, what use are our findings derived from our emphasis on construct modeling, measurement foundations, and digital learning? Put more specifically, what use is this research for writing program administrators (WPAs) in their work of programmatic course design, particularized writing center support, outreach of writing across the curriculum and writing in the disciplines, and outcomes assessment?

To begin, *WPAs should realize proposed curricular changes are best understood as goals to be approached though programmatic strategies*. The theory of change notes that for change to occur we must identify indicators (in this case, the students' perception of theirs and others' peer reviews) that might point to the preconditions that students have about themselves as reviewers and the reviews they receive. Identifying assumptions is key to beginning to effect change. It is only then that we can begin to intervene within the classroom to work towards the long-term goal of having students become more effective, critical writers.

Similarly, we might wish to consider theory-based evaluation, in particular when examining what assumptions exist within the program currently (Weiss, 1995). *A natural next step to implement the findings of our study would be for WPAs to systematically examine how evidence reacted to gender, ethnicity, and race is manifested within the classroom at their own institutions*. The students are coming into the first-year classroom with a range of high school experiences, in many ways related to complex backgrounds. How is the writing program filling these potential gaps? Similarly, are there gender and/or racial biases manifesting in the way the curriculum is taught? Not only are these sorts of questions imperative in order for generalizations to be made (and identified) when potentially applying these results to other writing programs,

this study has shown that they are prevalent when examined in terms of peer review. It is here we find the turn to social justice that has recently emerged in writing studies (Inoue, 2015; Kelly Riley & Whithaus, 2016; Poe & Inoue, 2016).

In addition, there is the practical response of the instructors who teach using the software of *MyR* and similar digital platforms. Here, it is important to recall that these are not digital filing cabinets; rather, they are construct-specific platforms designed as digital ecologies in which writing practices are fostered. Working in these environments is challenging and time-consuming. Instructors are expected to score students' peer reviews, as noted above, in terms of quality, quantity, and the tone of the review. Students are then reliant on the instructor to convey whether or not what they provided was a constructive, useful peer review. Instructors are encouraged to address peer review in the class beforehand (by modeling peer review) and then afterwards (as the students reflect on what comments they found useful). These activities require careful planning. *WPAs should therefore be mindful that teaching in platforms such as MyR is not a casual business; instruction in these construct-specific platforms requires attentive curricular design and collaborative planning when the curriculum is delivered across multiple sections of the same course*.

By extension, many of these classes also utilize individual student-teacher conferences that take place after the peer reviews have been received—work that extends beyond the *MyR* platform into the office and writing center. This extension gives the instructor the opportunity to directly address the student's writing, in another light, modeling what peer review comments might look like and giving the students a direct comparison to their received peer reviews. Bringing the clearly human part—in this case, classroom discussion and student-instructor conferences—to the digital practice of peer review aligns with the current implementation of digital humanities in the English higher education classroom, thus utilizing digitization to help better address English concerns, without eliminating current practices (Kirschenbaum, 2012). *As part of the larger planning effort, WPAs should be mindful that the introduction of a digital platform such as MyR does not end when the student turns from the screen; instead, the information gained from these systems permeates the entire writing program.*

As a safeguard against such a ubiquitous presence, there is a need for diverse perspectives. For example, Kristen Intemann (2010) observes that both feminist empiricism and standpoint feminism are useful in providing alternative stances when information is to be interpreted and used. In uncovering these differences in peer review, seemingly by the social markers of gender, race, and ethnicity, we must also recognize the importance of including the voices of diverse peoples in interpreting and implementing changes, in order for the change to be effective. It is key to remember that this discussion is bounded by social markers, which in the field of academia often ends up being the position of "insider," while many of these students are necessarily "outsiders" (Intemann, 2010), though these terms are inclusive of many other models, including gender, race, and socioeconomic status. *Related to social justice turn in writing studies research, inclusion of diverse perspectives in information interpretation and use is critical to the planning work of WPAs as they engage digital platforms*.

Such advice may not come as a surprise to WPAs interested in intrapersonal and interpersonal domains of the writing construct. If writing programs are to fulfill their goals of helping students become effective writers, WPAs know that taking student voices into account must become central to the change. Nevertheless, fostering these domains requires work of a different sort. Keeping peer review central to writing curriculum gives students a voice in their community of writers, but this desire for centrality must also be accompanied by modeling the exercise within the classroom. We cannot assume that students, even the ones who score well on our rubrics, are capable of effective, constructive peer reviews. We must also consider what long-term goals are going to help these students, based on their experiences with peer review. While it might be viewed as a stretch to assume if a student is ineffectual in their peer review they are also ineffectual in their writing, the research shows that offering constructive feedback makes writers more aware of what makes writing "good" (Jarratt et al., 2009; Meizlish et al., 2013). *By extension, WPAs will benefit by recognizing that pedagogies must be developed to support students in becoming effective peer reviewers—pedagogies that are related to, but distinct from, those targeted at exclusively cognitive dimensions of the writing construct*.

7.0 Directions for Further Research

Because of the nature of first-generation research, there are many limitations to this study. This study in many ways serves the purpose of merely opening the door to discussions on race and gender in relation to students' participation in the peer review process. From there, we must then connect peer review to the writing process. While our study scratches the surface of the relationship between students and peer review, our analysis does not include how these students went on to perform in their final drafts. Thus, a natural next step is to expand the study to include the students' full writing process. The instrumental value to the research provided in this article can be summarized in three areas of consideration that we believe are worthy of extended pursuit.

1. *Consider the advantages of expanded notions of the writing construct*. Gallagher (2016) has correctly observed that "writing behaviors (like all behaviors) are shaped by the social environments in which they are undertaken" (p. 258). In extending that observation, expanded ideas of the writing construct allow logical examination of the many facets of this most complex of human actions. Viewing peer review as integral to the construct—not as an external pedagogical principle, but as part of the behavior itself—transforms the solitary act of writing into one of communal investigation. It is the nature of that community that has, in essence, been narrowly investigated in the study reported here.

2. *Consider information analysis in terms of opportunity to learn*. In its most recent projections, the U.S. National Center for Educational Statistics (Hussar & Bailey, 2014) estimates that public school K-12 enrollments of Hispanic, Asians/Pacific Islander, and other students of two or more races are expected to dramatically rise. In terms of increase against the 2011 benchmark, there will be a 20% rise for students who are Asian/Pacific Islander, a 33% rise for students who are Hispanic, and a 44% rise for students who are two or more races. The

number of high school graduates is projected to reflect a 23% rise by 2022 for students who are Asian/Pacific Islander and a 64% rise by 2022 for students who are Hispanic. Reflecting these systemic shifts posed by the rapid demographic evolution in the U.S. from the 1970s to 2060, the Center for American Progress and the American Enterprise Institute project the race/ethnicity composition of the electorate to 2060 (Teixeira, Frey, & Griffin, 2015). The demographic changes they project are so significant—from 80% White citizens in 1980 to 44% in 2060—that they classify these shifts as superdiversification. During that same period, Hispanic citizens are projected to grow from 6% to 29% and Asian Americans from 2% to 15%. Anticipating that which is surely to come, radical shifts are required to ensure fairness for all students. Studies will be required to examine the relationship between expansion of the writing construct and increasing opportunity to learn.

3. *Consider digital ecologies as a way to advance writing instruction for all students*. White, Elliot, and Peckham (2015) have advanced the concept of ecology as fundamental in understanding the environments in which students write. As a science of communities, ecologies are distinct and, as such, information about them is both robust and limited. In the present study of peer review, for example, we have used focused concepts of peer review to reveal information about student learning; however, it would be a mistake to conclude that generalization inferences between, say, the study by Tucker (2014) and this one produced dissimilar findings. *MyR* provides a unique student experience, and generalization inferences will prove difficult across institutional settings. This drawback is compensated, however, by realizing that similar digital ecologies can produce results that can, in turn, allow extensive information to be provided in ways to improve site-specific student learning.

To realize the consequences of these three considerations, revisiting Figure 2 is in order. Look again at the first column of responses given by students who felt that they did not engage with the text of their peers. Now, return to Table 1 and recall that students were also given performance scores on their writing. If we compare the rubric scores of students who, upon self-reflection, gave themselves scores of 1 and compare those students to the majority of students who awarded themselves a score of 5, we realize just how inter-related the intrapersonal domain is to writing performance: Students who gave themselves the lowest self-reflection responses ($M = 2.82$, $SD = .82$) differed in their performance at statistically significant levels ($t(1214) = 3.56$, $p < .001$) from students who felt they had fully engaged with the text of classmates ($M = 3.08$, $SD = .69$). Put simply, there appears to be a relationship between feeling inadequate about peer review and subsequent poor writing performance.

Author Biographies

Ashley N. Reese is a Digital Teaching Fellow at the University of South Florida. Her Ph.D. is in children's literature from the University of Cambridge. Her current research focuses on peer review in the classroom, as well as gender and religion in turn-of-the-century, North American girls' books. She is the author of the forthcoming *The Rise of Girls' Literature* (Cambridge University Press).

Rajeev R. Rachamalla is an Application Project Manager at the University of South Florida. He received his master's in management information systems, with a concentration in data analytics and anti-money laundering, at USF. As a project manager, Rachamalla leads product development for the English Department's writing program, in addition to research to improve student writing skills. He is currently pursuing a second Master's in Business Administration at USF.

Alex Rudniy, **Ph.D**. is an assistant professor of computer science at the University of Scranton. He earned a doctorate in computer science from the New Jersey Institute of Technology. Dr. Rudniy was previously an assistant professor at Fairleigh Dickinson University and completed research sponsored by the National Science Foundation and National Security Agency.

Laura L. Aull is an associate professor of English and linguistics at Wake Forest University, USA. Her research focuses on linguistic analysis of student writing in higher education and can be found in journals in writing studies, applied linguistics, and writing assessment. She is the author of *First-Year University Writing: A Corpus-Based Study with Implications for Pedagogy*.

David Eubanks holds a Ph.D. in mathematics from Southern Illinois University and currently serves as Assistant Vice President for Assessment and Institutional Effectiveness at Furman University. He has worked on the practical side of assessing student learning, including student writing, since the late nineties. His research focuses on writing assessment, predictive analysis, and survey research, including the development of methods and software tools. Some of these applications can be found online at http://github.com/stanislavzza.

Acknowledgments

This research was performed under NSF Promoting Research and Innovation in Methodologies for Evaluation (PRIME) Program Award 1544239: Collaborative Research—The Role of Instructor and Peer Feedback in Improving the Cognitive, Interpersonal, and Intrapersonal Competencies of Student Writers in STEM Courses. Joe Moxley and Norbert Elliot contributed to the design of the research reported here.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Aull., L. (2015). *First-year university writing: A corpus-based study with implications for pedagogy*. New York: Palgrave Macmillan.

Behizadeh, N., & Engelhard. G. (2015). Valid writing assessment from the perspectives of the writing and measurement communities. *Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana*, *52*(2), 34-54.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–27). Hillsdale, NJ: Erlbaum.

Chang, C.Y-h. (2016). Two decades of research in L2 peer review. *Journal of Writing Research*, *8*(1), 81-117.

Cheng, W., & Warren, M. (1997). Having second thoughts: Student perceptions before and after a peer assessment exercise. *Studies in Higher Education*, *22*(2), 233–239.

College Board (2016). *2016 college-bound seniors: Total group profile*. New York, NY: College Board. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/sat/total-group-2016.pdf

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings?. *Assessing Writing*, *18*(1), 100-108.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, *24*(3), 331–350.

Elliot, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment*, *9*(1). Retrieved from http://journalofwritingassessment.org/article.php?article=98

Falakmasir, M. H., Ashley, K. D., Schunn, C. D., & Litman, D. J. (2014). Identifying thesis and conclusion statements in student essays to scaffold peer review. In S. Trausan-Matu, K.E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent Tutoring Systems. ITS 2014. Lecture notes in computer science, vol 8474.* Springer.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387.

Gallagher, C. W. (2016). What writers do: Behaviors, behaviorism, and writing studies. *College Composition and Communication*, 68(*2*), 238–265.

Gibson, A., Kitto, K., & Bruza, P. (2016). Towards the discovery of learner metacognition from reflective writing. *Journal of Learning Analytics*, *3*(2), 22–36.

Greeno, J. G., & Gresalfi, M. S. (2008). Opportunities to learn in practice and identity. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 170-199). Cambridge, UK: Cambridge University Press.

Hamer, J., Purchase, H., Luxton-Reilly, A., & Denny, P. (2015). A comparison of peer and tutor feedback. *Assessment and Evaluation in Higher Education*, *40*(1), 151–164.

Hart-Davidson, W., McLeod, M., Klerkx, C., & Wojcik. M.. (2010). A method for measuring helpfulness in online peer review. In *Proceedings of the 28th ACM International Conference on Design of Communication (SIGDOC '10)*. ACM, New York, NY, USA, 115–121.

Haswell, R.H. (2005). NCTE/CCCC's recent war on scholarship. *Written Communication*, *22*(2), 198–223.

Hayes, J. R. (2012). Modeling and remodeling the writing construct. *Written Communication*, *29*(3), 369–388.

Hewett, B. L. (2015). A review of WriteLab. *WLN: A Journal of Writing Center Scholarship*, *40* (3-4), 8–19.

Hu, G., & Lam, S.T.E. (2010). Issues of cultural appropriateness and pedagogical efficacy: Exploring peer review in a second language writing class. *Instructional Science*, *38*, 371–394.

Hussar, W. J., & Bailey, T. M. (2014). *Projections of education statistics to 2022* (NCES 2014-051). 41st ed. U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Retrieved from http://nces.ed.gov/pubs2014/2014051.pdf

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing for a socially just futur*e. Fort Collins, CO & Anderson, SC: WAC Clearinghouse and Parlor Press.

Intemann, K. (2010). 25 years of feminist empiricism and standpoint theory: Where are we now? *Hypatia*, *25*(4), 778–796.

Jarratt, S. C., Mack, K., Sartor, A., & Watson, S. E. (2009). Pedagogical memory: Writing, mapping, translating. *WPA: Writing Program Administration*, *33*(1– 2), 46– 73.

Kane, M. T. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 64–80). New York, NY: Routledge.

Kelly Riley, D., & Whithaus, C. (2016). A theory of ethics for writing assessment. [Special issue]. *Journal of Writing Assessment*, *9*(1). Retrieved from http://journalofwritingassessment.org/article.php?article=99

Kirschenbaum, M. (2012). Digital humanities as/is a tactical term. In M.K. Gold & L.F. Klein (Eds.), *Debates in the digital humanities* (n.p.). Minneapolis, MN: University of Minnesota Press.

Lawrence, S.M., & Sommers, E. (1996). From the park bench to the (writing) workshop table: Encouraging collaboration among inexperienced writers. *Teaching English in the Two-Year College*, *23*(2), 101–9.

Leijen, D.A.J. (2017). A novel approach to examine the impact of web-based peer review on the revisions of L2 writers. *Computers and Composition*, *43*, 35–54.

Leijten, M., Van Waes L., Schriver, K., & Hayes, J. R. (2014). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research*, *5*(3), 285–337.

Liang, M-Y. (2010). Using synchronous online peer response groups in EFL writing: Revision-related discourse. *Language Learning & Technology*, *14*(1), 45–64.

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, *18*, 30–43.

McLaughlin, P., & Simpson, N. (2004). Peer assessment in first year university: How the students feel. *Studies in Educational Evaluation*, *30*(2), 135-149.

Meizlish, D., LaVaque- Manty, D., & Silver, N. (2013). Think like/write like. In R. Thompson (Ed.), *Changing the conversation about higher education* (p. 53–74). New York, NY: Rowman & Littlefield.

Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement*, 53(*3*), 265-292.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective*, *1*(1), 3–62.

Mongo (2016). *MongoDB Atlas best practices*. New York, NY: Mongo. Retrieved from https://www.mongodb.com/collateral/mongodb-atlas-best-practices

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn.* Cambridge, UK: Cambridge University Press.

Moxley, J. M., & Eubanks, D. (2016). On keeping score: Instructors' vs. students' rubric ratings of 46,689 essays. *Writing Program Administration*, *39*(2), 53–80.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Committee on Defining Deeper Learning and 21st Century Skills, J. W. Pellegrino & M. L. Hilton, Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education (Eds.). Washington, DC: The National Academies Press.

National Research Council. (2013). *Frontiers in massive data analysis*. Washington, D.C.: The National Academies Press.

Nystrand, M. (1984). Learning to write by talking about writing: A summary of research on intensive peer review in expository writing at the University of Wisconsin—Madison. ED 255 914. Retrieved from https://eric.ed.gov/?id=ED255914

Paulus, T.M. (1999). The effect of peer and teacher feedback on student writing. *Journal of Second Language Writing*, *8*, 265–289.

Poe, M., & Inoue, A. B. (2016). Writing assessment as social justice [Special issue]. *College English*, *79*(2).

Raymond, R.C. (1989). Teaching students to revise: Theories and practice. *Teaching English in the Two-Year College*, *16*(1), 49–58.

Ross, V., Liberman, M., Ngo, L., & LeGrand, R. (2016). Weighted log-odds-ratio, informative dirichlet prior method to enhance peer review feedback for low- and high-scoring college students in a required first-year writing program. *Proceedings of the EDM 2016 Workshop and Tutorial*. Retrieved from http://ceur-ws.org/Vol-1633/ws2-paper4.pdf

Rudniy, A., & Elliot, N. (2016). Collaborative review in writing analytics: N-gram analysis of instructor and student comments. *Proceedings of the EDM 2016 Workshops and Tutorials.* Raleigh, NC, USA, June 29, 2016, 1–8.

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test taker's ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology*, *34*(4), 665–693.

Struyven, K., Dochy, F., Janssens, S., & Gielen, S. (2006). On the dynamics of students' approaches to learning: The effects of the teaching/learning environment. *Learning and Instruction*, *16*(4), 279–294.

Teixeira, R., Frey, W. H., & Griffin, R. (2015). *States of change: The demographic evolution of the American electorate, 1974-2060*. Washington, DC: Center for American Progress, American Enterprise Institute, & Brookings Institution. Retrieved from https://cdn.americanprogress.org/wp-content/uploads/2015/02/SOC-report1.pdf

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*(3), 249-76.

Tsui, A. B. M., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing*, *9*(2), 147–170.

Tucker, R. (2014). Sex does not matter: Gender bias and gender differences in peer assessments to contributions to group work. *Assessment & Evaluation in Higher Education, 39*(3), 293–309.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. I. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), *New approaches to evaluation community initiatives: Concepts, methods, and contexts* (p. 65–92). New York, NY: The Aspen Initiative.

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation, 19*(1), 21–33.

Wen, M. L., & Tsai, C. C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education, 51*, 27–44.

White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. Logan, UT: Utah State University Press.

White, T. (2015). Hadoop: The definitive guide. (4th ed.). *Storage and analysis at internet scale*. Sebastopol, CA: O'Reilly Media.

Willey, K., & Gardner, A. (2010). Investigating the capacity of self and peer assessment activities to engage students and promote learning. *European Journal of Engineering Education, 35*(4), 429–443.

Wilson, M. J., Diao, M. M., & Huang, L. (2015). 'I'm not here to learn how to mark someone else's stuff': An investigation of online peer-to-peer review workshop tool. *Assessment & Evaluation in Higher Education, 40*(1), 15–32.