Slouching Toward Sustainability: Mixed Methods in the Direct Assessment of Student Writing

by Jeff Pruchnic, Chris Susak, Jared Grogan, Sarah Primeau, Joe Torok, Thomas Trimble, Tanina Foster, and Ellen Barton, Wayne State University

The development of present-day assessment culture in higher education has led to a disciplinary turn away from statistical definitions of reliability and validity in favor of methods argued to have more potential for positive curricular change. Such interest in redefining reliability and validity also may be inspired by the unsustainable demands that large-scale quantitative assessment would place on composition programs. In response to this dilemma, we tested a mixed-methods approach to writing assessment that combined large-scale quantitative assessment using thin-slice methods with targeted, smaller-scale qualitative assessment of selected student writing using rich features analysis. We suggest that such an approach will allow composition programs to (a) directly assess a representative sample of student writing with excellent reliability, (b) significantly reduce total assessment time, and (c) preserve the autonomy and contextualized quality of assessment sought in current definitions of validity.

Things fall apart.

--William Butler Yeats, *The Second Coming*, 1919

Writing program administrators often find themselves torn between devoting time and resources to pursuing the quantitative assessment of a large sample size of student writing or the qualitative assessment of a smaller sample size. The former approach is likely the evaluation that will be most convincing to external stakeholders, while the latter is often more likely to provide a richer basis for instituting the kinds of curricular changes that will improve instruction. Each alone can also be time-intensive and thus resource-consuming, a situation that, particularly in austere times, can leave faculty with a difficult decision to make. Indeed, this predicament provides an important context for scholarly debates over how to prioritize different types of validity in writing assessment: those based on traditional quantitative measures like interrater reliability and representative sample size or those based on qualitative measures that provide insight for curricular improvements. While they may be equally important in different situations, the time and resources needed to fulfill both types of validity are likely to exceed the resources of many programs.

In this article, we suggest one solution to this problem might be the integration of “thin slice” processes for scoring texts into our assessment methods. While they are a quantitative method more common to research in Behavioral Psychology, we document our success in using a thin slice process as part of a broader mixed-methods approach to writing assessment that combines large-scale quantitative assessment with targeted, smaller-scale qualitative assessment. In particular, we suggest that the efficiency of thin slice methods for quantitative scoring can allow large writing programs to reduce assessment time while also increasing the interrater reliability and sample size of their assessment process.

Literature Review

Reliability

The recent growth of assessment mandates and expectations for writing assessment have coincided with shifts in scholarly debates over the best ways to define or measure reliability in these processes. Indeed, the turn toward increased expectations for formal assessment of writing in higher education led almost immediately to a turn against statistical definitions of reliability and validity in assessment scholarship. When met with the pressures of what White, Elliot, and Peckham (2015) have called the “Age of Accountability” (p. 17) in writing assessment, many scholars and instructors took issue with the problematic decontextualization of program goals and student achievement that can occur in large-scale quantitative assessment. More specifically, writing assessment scholars came to increasingly question the dominant position of inter-rater reliability (IRR), a statistical measurement of the consistency of scores produced by multiple evaluators, in validating assessment practices and results.

An important early critique of IRR’s privileged position in writing assessment by Cherry and Meyer (1993) focused on two common mistakes that they suggested led to its outsized influence on assessment practices. First, there was the simple mistake of equating reliability with validity, that is, that proving reliability would guarantee validity automatically. As Cherry and Meyer (1993) reminded us, reliability is necessary for but not itself constitutive of validity: “A test cannot be valid unless it is reliable, but the opposite is not true: a test can be reliable but still not be valid” (p. 110). Second, Cherry and Meyer also suggested that while there may have been a problematically large emphasis on reliability in assessment scholarship, there was also a problematically small conception of different kinds of reliability as useful in the evaluation of assessment practices. More specifically, they detailed, some scholars seemed to presume IRR as the only reliability measurement worthy of consideration, but IRR is one of multiple types of reliability available to researchers. Such concerns about the presumed overemphasis on reliability, and IRR specifically, were soon met with more expansive and innovative reconsiderations of reliability. These approaches would go beyond correcting the conflation of reliability with validity to query whether validity measures needed to account for reliability at all. On the one hand, scholars increasingly came to answer in the affirmative to the titular question of Moss’s 1994 essay, “Can There Be Validity Without Reliability?” (O’Neill, 2011). On the other hand, however, there is certainly the need to validate assessment through the use of formal criteria that would overlap with reliability: As O'Neill (2011) wrote, while for many scholars of writing assessment "a purely quantitative, statistical approach to reliability does not fit well with what we value" ("Reframing Reliability," para. 7), many of the same individuals also "recognize the significance of reliability and that there are some positive, useful values that reliability supports, so we cannot dismiss it out of hand" ("Reframing Reliability," para. 7).

Validity

A number of efforts to reframe validity in writing assessment have coalesced around the privileging of the impact of an assessment practice on creating positive curricular change rather than its coherence with respect to IRR or other statistical methods for measuring consistency of evaluation. In many ways following a similar reaction to the push for large-scale writing assessment in K-12 education that led researchers to argue for "the importance of expanding the concept of validity to include explicit consideration of the consequences of assessment use" (Moss, 1992, p. 229), educational measurement scholars and writing assessment scholars in higher education were soon advocating not only for internal stakeholder control of program outcomes relevant to assessment, but also for assessment methods determined by those same stakeholders to have the best potential for curricular change, as opposed to methods requiring the achievement of statistical reliability and validity. For such scholars, validity is defined and determined not by the consistency of raters' results (IRR) but instead by how consequential a method proves to be in producing effective curricular change. In this redefinition, reliability is measured not by the degree to which an assessment method produces consistent rankings across readers, but rather by the degree to which a given method can form a "critical standard with which communities of knowledgeable stakeholders make important and valid decisions" (Huot, 1996, p. 558). More specifically, this conceptualization repositioned validity as both a framework and a consideration of the consequences of assessment. In contrast to statistical evidence indicating validity, defined as the match between the measuring instrument and the construct being measured, the new definition instead prioritized the social impact of measurement on persons or programs being evaluated and the reasonableness of using specific assessment results as warrants for particular actions (Kane, 2006; Messick, 1989; O'Neill, Moore, & Huot, 2009; White et al., 2015). This expanded definition of validity, once controversial, is now widely accepted across fields and is codified in the *Standards for Educational and Psychological Testing* (2014), collaboratively published by the American Psychological Association, American Educational Research Association, and National Council of Measurement in Education (White et al., 2015, p. 82). Both White et al. (2015) and Elliot (2015) emphasized the importance of Messick's (1989) contribution to this changing definition of validity. Messick (1989) called for a definition of validity as:

> an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences* and *actions* based on test scores or other modes of assessment… Broadly speaking, then, validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use. (p. 13)

As further evidence of the move toward consequences or uses of assessment results, scholars began using the term "validation" to distinguish between the process of evaluating the use and interpretations of an assessment and statistical definitions of validity (Elliot, 2015; White et al., 2015).

Elbow (2012) presented a concise and effective argument for the value of validation measures in his essay "Good Enough Evaluation: When it is Feasible and When is Evaluation Not Worth Having?" In that text, part of a collection honoring Ed White, Elbow adapted White's focus on "balancing" the pragmatic need for assessment with the possible dangers of inaccurate or damaging assessment results (when assessment is "not worth having"). In particular, Elbow worried over the ratio of danger versus value in quantitative methods in which "numbers" are used to justify the validity of an assessment design as well as to present the significance of its results. Elbow suggested even in cases when positive "numbers" are produced by large-scale quantitative assessment, these data are often not particularly useful to a program without additional curricular context or instructor insight. For Elbow, the problem is that quantitative measurements seem to satisfy stakeholders in upper administration but often fail to provide the curricular insight gleaned from qualitative approaches that shift available assessment time away from scoring and toward discussion. While he did not present a specific solution to this conundrum, Elbow suggested programs can concentrate on using valid assessment methods for the purposes of identifying useful trends of assessed criteria in larger groups of sample student portfolios. With this approach,

> Programmatic evaluations could validly identify writers whose before and after portfolios show their degree of improvement near the top of what can be expected—and also those whose degree of non-improvement puts them at the bottom. These more trustworthy single numbers would be suggestive and useful, even though they speak of only a minority of students. (Elbow, 2012, p. 320)

However, for some, Elbow's solution might lead assessment personnel to too quickly abandon quantitative measures that may be valuable to programs and may also be preferred by external stakeholders. Richard Haswell (2012) addressed both of these points in his chapter in the same collection, “Fighting Number with Number.” Haswell traced what he identified as an outright fear of quantitative data in recent assessment scholarship in Writing Studies: “Numbers are like microbes and fires—people both need and fear them” (p. 413). But in Writing Studies, the contradiction takes on an added irony, leading Haswell (2012) to question, “Why shouldn’t our profession, which studies and teaches the way language does dubious business with art and truth, buy into numbering as a profitable trade in persuasion and argumentation?” (pp. 413-414). He presented five case studies in which a program's generation of quantitative assessment data successfully helped programs to respond to criticisms or aggressive inquiries about student learning from external. Thus, despite the many positive outcomes of spending the majority of assessment efforts on qualitative interpretation and curricular reform as opposed to quantitative assessment, Haswell (2012) suggested this practice might also leave the same programs vulnerable to external stakeholders’ critiques of the reliability of their assessment results and/or with little defense against assessments performed by external stakeholders that make use of quantitative data. When quantitative assessment and the fulfillment of statistical conditions for IRR or a representative sample size are absent, program coordinators cannot defend themselves through recourse to quantitative data, or, in Haswell’s (2012) colorful words, they no longer have the ability to “fight numbers with numbers” (p. 414).

Taken together, Haswell (2012) and Elbow (2012) identified the often-conflicting stakeholders that have to be considered in assessment design. Elbow, in accordance with the influential critiques of psychometric assessment practices made by scholars like Moss (1992, 1994) and Huot (1996), emphasized how writing instructors and program directors will benefit more from contextualized qualitative assessment of student writing, even if that assessment is modest in regard to its sample size and fails to meet (or ignores entirely) reliability in scoring. Haswell drew attention to the ways in which neglecting quantitative measurements and factors like IRR and representative sample sizes might undermine the autonomy of a program’s assessment efforts insofar as these factors are often important, perhaps essential, to members of an institution’s upper administration as well as external stakeholders that can impact a university’s standing and resources.

An obvious answer to this dilemma would be to have the best of both worlds and balance quantitative assessment with validity factors like IRR and the use of representative sample size with validation factors more likely to directly lead to the improvement of writing curricula. However, it is often time-consuming (and thus resource intensive) to conduct assessment meeting statistics-driven reliability and validity factors by itself, and to combine it with qualitative assessment measures geared towards producing concrete curricular reform would, of course, only increase the necessary time and labor needed for assessment. Indeed, in addition to concerns regarding social impact and context of assessment, the move toward validation approaches to assessment may itself have been additionally inspired by the seemingly unsustainable demands that large-scale quantitative assessment would place on (often understaffed and resource-strapped) composition programs.

Historically, innovation in writing assessment methods has been affected by the desire to reduce strain on time and program resources. Though many Writing Studies scholars rightly defend the value of portfolio assessment, citing the wealth of context-rich data gleaned from assessing student writing portfolios (Condon, 2011), it has become increasingly difficult to reconcile a commitment to authentic assessment of writing portfolios on the one hand and, on the other, the need to generate datasets based on statistically representative sample sizes at large public research universities. Ed White’s (2005) Phase 2 Portfolio Assessment is often celebrated as “a means for scoring portfolios that […] allows for relatively efficient grading where portfolio scores are needed and where time and money are in short supply” (p. 583). However, even this streamlined process still leaves WPAs at larger institutions to somehow balance the sum time and cost of training readers, scoring hundreds of portfolios, analyzing data, and presenting valid arguments about student outcomes with less resources, short timelines, and high stakes. In the current assessment culture, “the financial burden of this method is too great because of the investment in time and human resources” (Behizadeh, 2014, p. 6) required to generate data and make the sorts of claims about student outcomes that administrators and legislators expect.

It was also, of course, the markedly time-intensive process of direct writing assessment that inspired the trend toward standardized testing of writing skills in mid-twentieth century educational writing assessment (Elliot, Plata, & Zelhart, 1990, pp. 35-40); it is therefore not at all surprising that writing scholars near the end of the century would be alarmed when faced with the prospect of assessing a representative sample of student writing, particularly when they were also expected to achieve statistical IRR and validity. Thus, the move toward redefining validity to emphasize an assessment’s potential for positive curricular change often intersected with concerns about the sustainability of writing assessment methods that would satisfy more homodox definitions of validity. Indeed, these two concerns are connected: The more time one spends attempting to perform quantitative assessment at the size and scope that would satisfy statistical reliability and validity, the less time, it seems, one would have to spend determining and implementing the curricular practices that would support the learning that instructors truly value.

Thin-Slice Methods

Responding to calls for mixed methods approaches in the assessment literature, particularly White et al. (2015), we found our approach to the direct assessment of a representative sample of student writing in an unlikely place—a quantitative approach from Behavioral Psychology called thin-slice methods (Ambady, Bernieri, & Richeson, 2000). (Thin-slice methods were popularized as “fast cognition” in Gladwell’s 2005 best-seller *Blink*.) Originally developed to decrease coding time and burden in observational studies of lengthy face-to-face social and institutional interactions, thin-slice methods select and assess relatively small representative “slices” of longer interactions for multiple raters to score with a common instrument. In comparison to full observational coding of entire interactions, such as medical appointments or teaching sessions, thin-slice methods have proven to be surprisingly reliable. For example, raters could predict end-of-term teacher evaluation scores by assessing and scoring a set of three 10-second silent video slices of a teaching session as well as raters assessing and scoring the entire teaching session (Ambady & Rosenthal, 1993). Similarly, raters could predict the incidence of surgery malpractice claims by assessing and scoring two 10-second audio clips from the beginning and end of a medical appointment as well as raters who assessed and scored the full appointment (Ambady, LaPlante, Nguyen, Rosenthal, Chaumerton, & Levinson, 2002).

Thin-slice methods have been used to support research in multiple and diverse domains such as education (Kulik, 2001); marketing (Ambady, Krabbenhoft, & Hogan, 2006); computer science (Stecher & Counts, 2008); medicine and behavioral health (Ambady, Koo, Rosenthal, & Winograd, 2002; Carcone et al., 2015; Henry & Eggly, 2013), and multiple branches of psychology (Grahe & Bernieri, 1999; Lippa & Dietz, 2000; Murphy, 2005; Murphy, Hall, & Colvin, 2003; Peracchio & Luna, 2006). Many thin-slice studies have been based on written transcripts of social interactions; recently, however, thin-slice researchers have begun to examine written language directly as data. For example, Stecher and Counts (2008) examined online social media profiles, finding that raters’ impressions of thin-sliced profiles reliably predicted raters’ impressions of full profiles. Our research is the first thin-slice study to investigate written language in an educational context and also the first study to apply and test thin-slice methods within the domain of assessment in Writing Studies.

We offer an intervention into the dilemma that seeks to combine the best elements of quantitative and qualitative assessment methods, and of statistical definitions of reliability and validity—and to do so in a time-efficient (and thus sustainable) manner. More specifically, we report here the results of an experiment integrating “thin slice” approaches for scoring texts quickly as part of a broader mixed-methods approach to writing. Through leveraging a thin slice approach, we were able to achieve excellent IRR in a large-scale direct assessment of student writing while significantly reducing assessment time. We were then able to use those results to anchor qualitative assessment driven toward curricular reform. We suggest that such a mixed methods approach allows writing programs to satisfy the demands of present-day assessment culture while maintaining the autonomy and contextualized quality of assessment sought in current definitions of validity and to do so in a resource-conscious manner.

Methods

For the field of Writing Studies, the affordances of thin-slice methods offer the possibilities of fully representative sampling and statistical measurements of reliability and validity, thereby providing a method to achieve high quality and sustainable large-scale direct assessment of student writing, when that is warranted in a particular assessment context. To test these affordances, we designed a mixed-methods study comparing the results of raters scoring thin-sliced versions of students’ end-of-semester reflective essays in FYC with raters scoring full versions of the same reflective essays.

Research Questions

Our quantitative and qualitative research questions (RQs) were the following:

1. In scoring the full reflective essays, what was the IRR of the Regular Team?
2. In scoring the thin-sliced reflective essays, what was the IRR of the Research Team?
3. What was the correlation of scores between the Regular and Research Teams?
4. What were the scoring times (by teams and by readers) for the Regular and Research Teams?
5. What kinds of textual features characterized reflective essays scored as sufficient (rubric categories 6, 5, 4) or insufficient (rubric categories 3, 2, 1) with respect to the judgment that a student writer had achieved or not achieved the Reflection outcome?

Study Site

This research took place at Wayne State University (WSU), a large urban public research university, with 27,500 students in over 380 degree programs in 13 schools and colleges. Approximately 18,000 students at WSU are undergraduates, many of them first generation college students and most of them working full- or part-time. The student body is 54% White and 36% racial/ethnic minority, with Black or African American students making up 21% of the total student body. At WSU, 64% of undergraduates attend college full-time, but 36% do not, a significant difference from flagship or regional universities. In 2014, WSU’s retention rate for first-to-second-year full-time students was 76%, much lower than the 83.5% average retention rate across peer institutions. Also, WSU’s 6-year graduation rate of 34.3% was well below the 59.25% average 6-year graduate rate across its peer institutions (Office of Budget, Planning, and Analysis, n.d.).

The Composition Program at WSU is located in the English Department. The first-year sequence features two courses: a basic writing course for students with ACT English scores of 20 and below (ENG 1010), and a traditional first-year composition (FYC) course for students with ACT English scores of 21 and above (ENG 1020). Approximately 65-70% of all freshmen place into ENG

1020, and every fall semester, around 1,200 students enroll in the course across approximately 65 sections. The FYC course has a common syllabus featuring standardized learning outcomes across four knowledge and practice domains: reading, writing, researching, and reflecting. These outcomes anchor an assignment sequence consisting of projects in rhetorical analysis, research-based argumentation, and visual argumentation. The course's current pass rate averages around 75%, and recent institutional research by the University has established the importance of passing FYC to student retention into the second year and forward to graduation in six years.

Approximately 58% of FYC sections at WSU are taught by graduate teaching assistants, with approximately 23% taught by full-time faculty, primarily lecturers, and 19% by part-time contingent faculty. Teaching assistants are trained to teach the course's common syllabus in a pedagogical practicum course taken during the first two semesters of their assistantship. Part-time faculty may audit the practicum course but are not required to attend. Both part-time instructors and teaching assistants are required, however, to attend a full-day teaching orientation at the beginning of each fall semester and must attend at least three hour-long teaching workshops held throughout the academic year. Both the fall orientation and academic year workshops are designed and facilitated by full-time lecturers, tenure-track faculty, and advanced part-time faculty and graduate teaching assistants who designed the course's common syllabus and assignment sequence.

The final student task in FYC's assignment sequence is a reflective argument essay based on White's (2005) Phase 2 assessment model. Since 2010, the Composition Program has used a version of White's end-of-semester essay as the primary assessment instrument for our first-year writing course and our two intermediate writing courses. The primary artifacts of Phase 2 assessment are the end-of-semester reflective essay and a traditional portfolio featuring a range of written products. We chose White's system for assessment based on his argument that the Phase 2 design has two important benefits for sustainable assessment. First, by focusing raters' attention on the shorter four- to six-page reflective essay, the model reduces the amount of time required to review and score student portfolios, which often run 30 to 40 pages in length. Second, because the reflective essay asks students to cite work within their portfolio as evidence of their achievements, raters can use their reading of essays to learn about the overall effectiveness of the curriculum and instructional approach (White, 2005).

Our use of the Phase 2 assessment model for FYC previously consisted of a reading and scoring activity held at the end of each semester in which a group of around 10 experienced instructors worked in pairs to read and score a randomly selected sample of end-of-semester essays, using a scaled rubric grounded in each of our FYC learning outcomes. Consistent with White's (2005) model and his description of other programs that use the model, our reading pairs used consensus scoring, first scoring each paper individually and then negotiating a final score. In cases where consensus could not be reached, a third reader scored the essay in question and then the average of all three scores was calculated to determine the essay's final score. Scoring data for the entire sample were then forwarded to the Director of Composition for distribution to other committees and administrators within the Composition Program.

Our Program's adoption of the Phase 2 model has fostered engagement with writing assessment within our department and garnered recognition from our University's administration, which, as many readers will recognize, is increasingly impacted by the culture of evaluation and assessment across higher education. However, despite these positive developments, our assessment program faced sustainability issues and methodological concerns. Portfolio assessment is labor-intensive and time-consuming. In fact, the Assessment Committee was aware that they had, over time, pragmatically chosen to assess less student writing, first eliminating reading material from the portfolio in 2011 and 2012, and then suggesting in 2015 that we use the reflective essay as the single artifact of assessment. Even reading the reflective essay alone still posed methodological problems in that it is hard to scale; it is difficult to evaluate a representative, randomized sample of student writing across all sections of the FYC course, and thus, hard to provide evidence meeting standards of validity and reliability in Phase 2 scoring. As a result, we have not been able to ask and answer important programmatic questions about, for example, the efficacy of our current curriculum, whether and how to revise it, and how to execute more targeted assessment of student writing in FYC. Even more problematically, we were making decisions about curricular and other matters that were not based upon a solid understanding of the writing of our entire student body because it was not based on a representative sample. Further, we found it increasingly difficult to practically sustain our efforts, even after abandoning full portfolio reading in favor of directly assessing students' reflective essay introductions to their portfolio. Just as importantly, we have also found it difficult to maintain the methodological validity of our current adaptation of the Phase 2 model and its potential to produce meaningful, data-based curricular improvement and professional development. Since assessment of FYC began in 2010, we estimate that each of our adaptations of Phase 2 scoring allowed us to assess writing from only 6-12% of the total course enrollments, which is far from a representative sample (around 26%) of all students finishing FYC.

We thus sought alternative methods for the direct assessment of student writing, methods that would allow us to assess a representative sample of student writing and that were sustainable in the context of a Composition Program with limited resources.

Data Collection

Teams. In previous FYC assessments, our Composition Program used 10-person teams. For this assessment of FYC reflective essays in Fall 2015, we recruited members to form two 10-person teams: the Regular Team, which would use Phase 2 assessment methods, and the Research Team, which would use thin-slice assessment methods. Unfortunately, two members of the Regular Team dropped out unexpectedly, so we conducted the assessment with an 8-member Regular Team and a 10-member Research Team. This must stand as a limitation of our study, although it does reflect the realities of conducting assessment in context.

Both teams were made up of experienced composition instructors, with similar breakdowns for rank. The Regular Team included three full-time faculty, two part-time faculty, and three graduate teaching assistants; the Research Team included four full-time faculty, three part-time faculty, and three graduate teaching assistants.

Representative sample. In Fall 2015, 1,377 students received a grade in FYC. For this assessment, we included only class sections that used the common reflective essay assignment. (Six sections of FYC were excluded because the instructors were piloting a different reflective assignment.) Across sections using the common reflective essay assignment, students submitted to instructors a total of 1,174 reflective essays. Using a sampling calculator, we determined that a representative sample of 1,174 essays was 290 (National Statistical Service, n.d.). We asked instructors to submit for assessment a random sample of six reflective essays (i.e., essays #4, 9, 12, 13, 14, and 19 from the alphabetical roster of each section). Our randomly selected representative sample consisted of 291 essays.

Materials. To develop a rubric for scoring our sample of reflective essays, we turned to the standardized reflective essay assignment in our FYC common syllabus, which ensured we were assessing students based upon what we actually asked them to write (see Appendix A). In the common syllabus for all sections of FYC in our Composition Program, the learning outcome for reflection was written with three key terms: “Use written reflection to *plan*, *monitor*, and *evaluate* one’s own learning and writing.” Two of these terms—*plan* and *monitor*—were metacognitive terms not mentioned in the reflective essay assignment prompt. Instead, the standardized assignment sheet for the reflective essay focused exclusively on the evaluative component of the learning outcome:

> Make an argument that analyzes your work in ENG 1020 in relationship to the course learning outcomes listed on the syllabus for the course. The body paragraphs of your essay should develop your main claim with evidence from your major works and experiences in this course.

Though planning and monitoring are important parts of other assignments in the course, those aspects of the reflection outcome were not assessed here because they are not described in the reflective essay assignment.

For our rubric (see Appendix B), we followed the emphasis of the assignment prompt in the description of the learning outcome to be assessed: “Use written reflection to evaluate one’s own learning and writing.” We also followed the assignment prompt in selecting and defining the two traits of our rubric: *argument*, defined as thesis, claim, relation to course outcomes; and *evidence*, defined as examples, analysis, experiences, discussion. On the advice of our statistician, we used a six-point Likert scale in the rubric for two reasons (Chang, 1994). First, the advantage of the six-point Likert scale is that it increases reliability when raters are knowledgeable in the domain of the study by offering categories for essays judged to be at the extreme ends of the continuum of the scale, thereby guarding against category inflation and deflation. Second, the specific assessment question in this study was whether a reflective essay demonstrated the student writer’s achievement of the Reflection learning outcome. A six-point Likert scale forced raters to make that judgment—was a student’s reflective essay sufficient to determine his/her achievement of the learning outcome (categories 6, 5, 4) or not (categories 3, 2, 1)—without allowing the rater to be undecided or neutral. Since all raters were experienced composition instructors, we expected them to be able to score using all points on a 6-point scale and to make an overall judgment in an assessment context.

Both the Regular Team and the Research Team used this common rubric for scoring their samples of reflective essays.

Procedures. All members of the Regular and Research Teams attended a 1-hour norming session on the first day of assessment, led by our Coordinator of the Assessment Committee in the Composition Program, who used the norming process previously used in regular assessment. The norming was based on two randomly selected reflective essays. All members of both teams read the full papers and gave them a score using the rubric. The scores were recorded on a white board, and the Coordinator then led a group discussion of why readers gave the scores they did with respect to the rubric. Both teams also attended a half-hour norming session on the second day of assessment, again led by the Coordinator. The process for the second norming was the same, but only one full essay was read, scored, and discussed.

Regular team assessment methods. Following the principles of White’s (2005) Phase 2 assessment, the Regular Team conducted paired readings and consensus scoring of the reflective essays as described by Haefner (2011): Two readers read the entire essay, used the rubric to score it individually, discussed the essays and their scores, and then came to a consensus score for each essay, with a third reading if necessary.

In this process, each essay was read and scored by only one two-member team. For this study, however, every fifth essay (20% of the essays) was read and scored by a second team in order to measure the IRR of the Regular Team (see Data Analysis below). Double-coding 20% of the data is routine for calculating IRR in the sciences and social sciences, a practice also used in observational studies in education, such as the well-regarded Teachstone Classroom Assessment Scoring System (CLASS) for PK-12 developed at the University of Virginia: "Research groups are often required to double-code 20% … to prove that they are reliable" (Vitiello, 2016, "Double Code," para. 1).

***Research team assessment methods**.* In describing the principles of thin-slice methodology, Ambady et al. (2000) suggested if observational data come from an interaction that has a definite beginning, middle, and end, the slices should come from these segments. We thus selected our thin slices from the beginning, middle, and end of the reflective essays, which happen to be the traditional categories of essay structure: the first paragraph (introduction), one paragraph from the middle page of the essay (body paragraph), and the final paragraph (conclusion). The first full paragraph on the middle page of the paper (e.g., page 3 of a five-page paper) was excerpted for the body paragraph. When an essay contained two middle pages, the second was used as the middle page (e.g., page 4 of a six-page essay). Ambady et al. (2000) did not identify a specific number of raters for any given thin-slice study: The studies they reported used a wide range of raters, from three to 193. Speaking methodologically and pragmatically, our study statistician noted the number of raters depends upon both the research questions of the study and the resources available for scoring. To compare the scoring times of the Regular Team and the Research Team, we designed our study to have two 10-member teams. In what we deemed to be the most efficient deployment of the 10 members of the Research Team, we divided the Research Team into two sub-teams of five, each scoring roughly half of the reflective essays: Five members scored 145 reflective essays, and five members scored 146 essays.

The members of the Research Team were given the title of the essay (if present) and the thin-slice paragraphs only for assessment, without access to the rest of the essay, and members of the Research Team did not consult each other during the readings. The Research Team read the thin-sliced reflective essays and scored them individually using the rubric. The final score for the essay was the average of the five raters' scores.

Data Analysis

Quantitative Analysis. To answer RQs #1 and #2, we measured IRR of the Regular Team and the Research Team using the Intra-Class Correlation Coefficient (ICC) (Hallgren, 2012). IRR measures the degree of agreement among raters' scores (the covariance of scores). We chose to use the ICC as our measure of reliability because it is an inferential measure used for interval (numerical) data scored by multiple raters. An ICC measurement will be high when there is little variation between raters' scores, and low when there is greater variation between raters' scores (Figure 1).

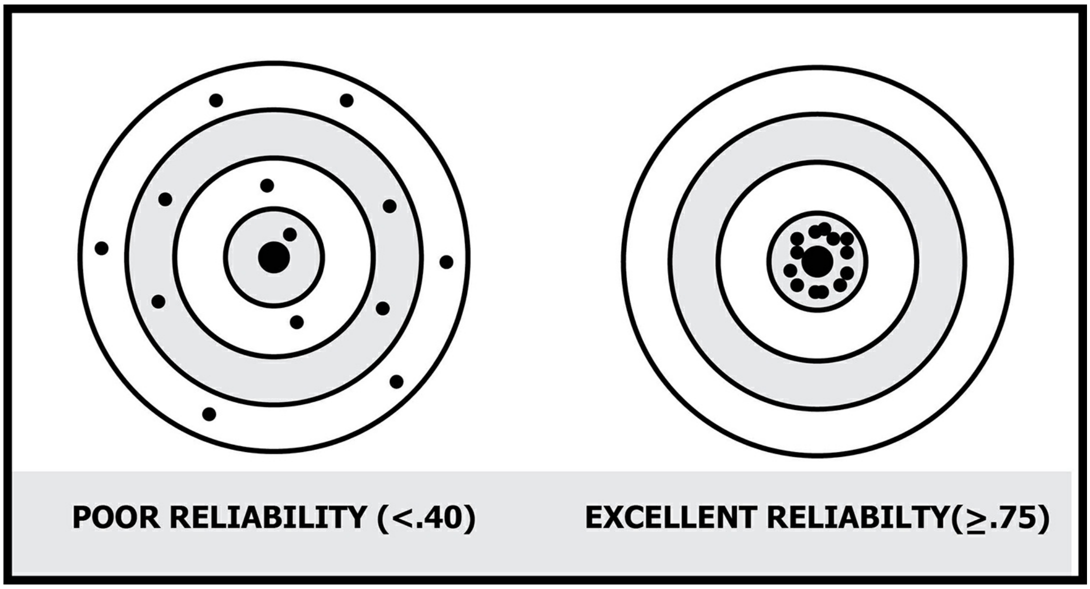

Figure 1 *Reliability*

To interpret the ICC results, we followed Cicchetti (1994): excellent reliability (>.75), good reliability (.60-.74), fair reliability (.40-.59), and poor reliability (< .40). The ICC thus indicates whether the members of the Regular Team and the Research Team were using the rubric to score the reflection essays consistently: In other words, ICC results indicate to what degree these scores would be reproducible given the same data, rubric, and conditions (Hallgren, 2012).

To answer RQ #3, we determined the correlation between the Regular and Research Teams using the Pearson Correlation Coefficient—Pearson's *r*—(Hinkle, Wiersma, & Jurs, 2003). We chose to use Pearson's *r* as our measure of correlation because our analysis used interval (numeric) data and because it is an inferential measure of similarity (a linear relationship on a line graph) between the scores of the two teams (see Figure 2).

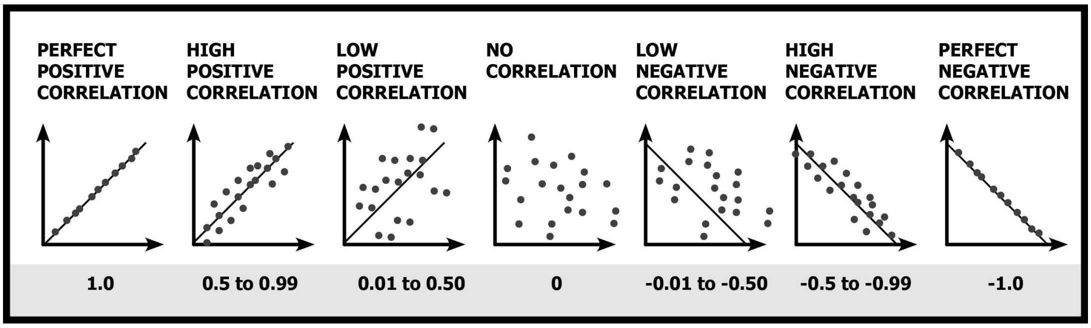

Figure 2 *Correlation*

While ICC measures the degree of variance in agreement (consistency), Pearson's *r* is a measure of similarity or the extent to which two sets of scores co-vary together. For example, if one rater scored in a series 1-2-3, and another rater scored in a series 2-3-4, they would be similar in that the series is the same for both raters: The lowest score (1 for the first rater or 2 for the second rater) increasing by one to the highest score (3 for the first rater or 4 for the second rater).

To interpret the results of the correlation calculation, we followed Mukaka (2012, p. 71):

- Perfect Positive Correlation +1.0
- High positive correlation +0.51 to +0.99
- Low positive correlation +0.01 to +0.50
- No correlation 0
- Low negative correlation -0.01 to -0.50
- High negative correlation -0.51 to -0.99
- Perfect Negative Correlation -1.0

The interpretation of the correlation (*r)* indicates whether the Regular Team and the Research Team were scoring the reflective essays similarly, although correlation findings must always be treated with caution: Here, a similar relationship does not mean an identical relationship, nor, as the well-known saying warns, does a correlational relationship imply causality.

To answer RQ #4, we recorded time information (hours and minutes) for both teams and all team members. We then compared the overall scoring time for each team as well as the average scoring times for team members.

Together, the ICC findings for reliability and the Pearson's *r* correlation findings for similarity provide evidence to determine whether the Regular Team and the Research Team were reading, assessing, and scoring the reflective essays in the same ways, that is, both consistently and similarly (RQs #1-3). The time information provides evidence to determine whether the Research Team coded more efficiently than the Regular Team, or not (RQ #4).

Qualitative Analysis. To answer RQ #5, we first ran a frequency analysis of the number of Research Team scores in each of the rubric categories (6-1, highest to lowest). We then randomly selected a set of 16 reflective essays from our sample, four from each of the rubric categories: Poor, Limited, Adequate, and Good. Because few essays were scored Excellent (*n* = 3) and None (*n* = 2), we did not select essays from these categories, although having scores in these end categories does indicate that raters were using all categories on the rubric's Likert scale as discussed above.

To analyze the reflective essays, we used rich feature analysis—a method of qualitative discourse analysis developed for Writing Studies (Barton, 2004). Rich feature analysis inductively or deductively looks for textual features that point to the relation between a text and its context. Rich features have both linguistic integrity (i.e., they are structural features of language and discourse, so they can be defined in linguistic terms and then categorized, coded, counted, and otherwise analyzed empirically) with contextual value (i.e., they can be conventionally connected to matters of function, meaning, analysis, interpretation, and significance). Meaning arises in part out of the repetitive and patterned use of rich features; if a feature is repeated within and across texts, it is likely to be typified and conventionalized as to form and function, and these conventional relations between features, patterns, and meanings describe the ways that rich features both reflect and shape the context of the text.

For our qualitative analysis, the members of the Research Team read the 16 essays holistically and listed textual features they noticed in their reading. No effort was made to focus team members' readings on the reflection outcome, the definitions of the rubric categories, the traits of the rubric, or any other deductive schema. In group sessions, members simply called out textual features they noticed from their readings. We then coded and categorized the features inductively in order to compare the differences between the rich feature profiles of essays scored as sufficient (6, 5, 4) or insufficient (3, 2, 1) with respect to the judgment that a student writer had achieved or not achieved the Reflection outcome of our FYC course.

Results

Quantitative Findings (RQs #1-4)

To answer RQs #1-2, we first compared the reliability of the Regular Team (Phase 2 methods) and the reliability of the Research Team (thin-slice methods) using the ICC measure (see Table 1).

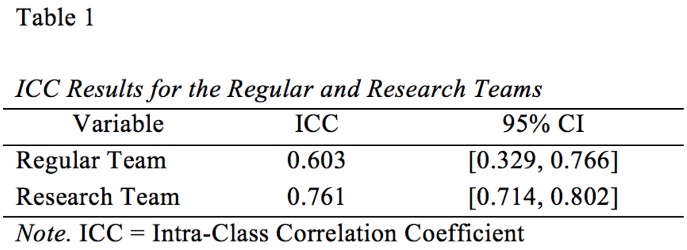

Table 1

ICC Results for the Regular and Research Teams

Variable	ICC	95% CI
Regular Team	0.603	[0.329, 0.766]
Research Team	0.761	[0.714, 0.802]

Note. ICC = Intra-Class Correlation Coefficient

Table 1 *ICC Results for the Regular and Research Teams*

Following Cicchetti (1994), our ICC results were .60 for the Regular Team (good reliability) and .76 for the Research Team (excellent reliability). Notably, the ICC for both teams was at the low end of their classifications: The classification of excellent reliability begins at 0.75, and the category classification of good reliability begins at 0.60, so the reliability of the Research Team (.76) was one full classification higher than the Regular Team (.60). Also of note, the upper and lower bounds of the 95% confidence level varied considerably across the two teams, which has implications for expected reliability in a replication of this study (again, given the same data, rubric, and conditions). For the highly reliable Research Team, the tight bounds indicate that the expected range of scores would be within the ICC classifications of good (.60-.74) or excellent (≥ .75) 95% of the time. For the less reliable Regular Team, however, the much wider bounds indicate that the expected range of scores could be within the entire set of IRR/ICC classifications, from poor (< .40) all the way to excellent (≥ .75).

To answer RQ #3, we then calculated the correlation between the Regular Team and the Research Team using the Pearson's *r* (see Figure 3).

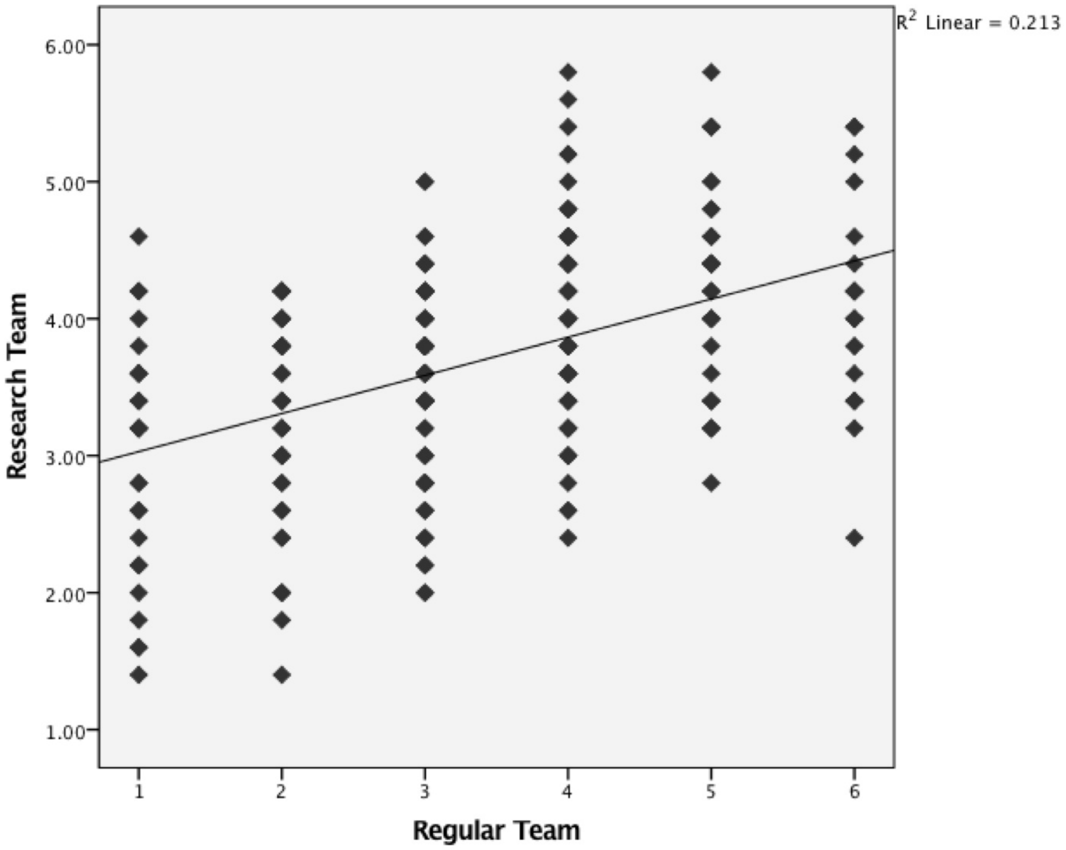

Figure 3 *Pearson's* r *for the Regular and Research Teams*

This graph depicts the correlation of the Regular Team scores, which assigned discrete numerical scores (x-axis) and the Research Team scores, which assigned an average of scores (y-axis). The line of best fit on the graph shows generally that the Regular Team and the Research Team scored similarly.

Following Mukaka (2012), there was a statistically significant correlation between the two teams: *r* = .462, low positive, but clearly trending toward high positive ($\geq$.51), indicating that as the Regular Team's scores increased, the Research Team's scores also increased. As noted above, this correlational result must be considered with caution: Though trending toward high positive, the correlation was low positive; however, it was a statistically significant correlation providing evidence that the members of the two teams were scoring similarly given the same essays, rubric, and conditions.

Another important finding from the quantitative component of our study was the time differential between the two teams (RQ #4). To score the entire set of reflective essays (see Figure 4), the Research Team (3,203 minutes or 53 hours and 23 minutes) spent a little more than half the time of the Regular Team (5,640 minutes or 94 hours).

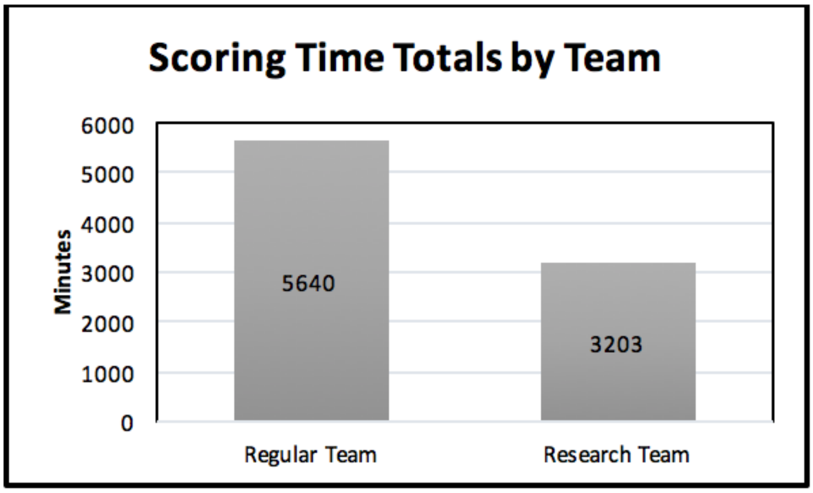

Figure 4 *Overall Scoring Time Totals by Team*

Figure 4 *Overall Scoring Time Totals by Team*

Similarly, the average scoring time of the members of the Research Team (320 minutes or 5 hours and 20 minutes) was a little less than one half of the average scoring time of the members of the Regular Team (705 minutes or 11 hours and 45 minutes) (see Figure 5). Not included in these calculations is the time a graduate student research assistant spent preparing the essays, which totaled approximately 11 hours. We did not factor this additional time into our comparison because it included tasks that were specific to the preparation of data for a research study rather than a typical assessment reading or that would be performed by instructors if thin-slicing was our standard assessment method (e.g. duplicating essays for a comparison study and anonymizing essays to meet the requirements of our institutional review board.)

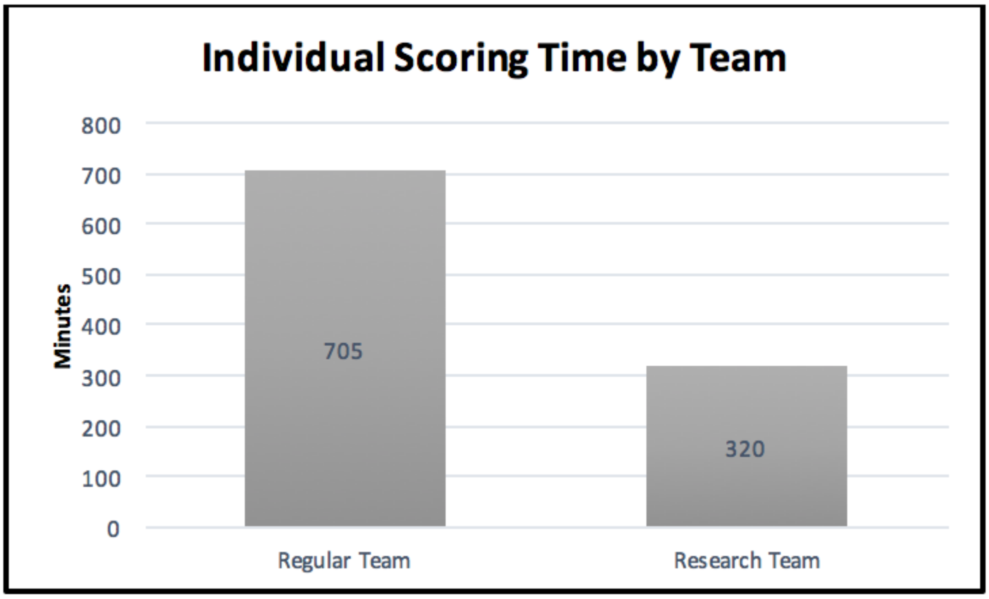

Figure 5 *Average Individual Scoring Time by Team*

Figure 5 *Average Individual Scoring Time by Team*

As would be expected, the use of thin-slice methods provided a considerable time savings for the direct assessment of reflective essays.

The conclusion of the quantitative component of the study was that both Phase 2 methods and thin-slice methods can reliably be used for a large-scale direct assessment of a representative sample of reflective essays written to demonstrate student writers' achievement of our FYC Reflection outcome: “Use written reflection to evaluate one's own learning and writing.” In fact, however, the Research Team using thin-slice methods was more statistically reliable in its scoring than the Regular Team using Phase 2 methods, and it was significantly more efficient as well. These findings are consistent with the thin-slice literature (Ambady & Rosenthal, 1993; Ambady et al., 2000), which has regularly found that raters using thin-slice methods can be as reliable, or even more reliable, than raters using full-slice methods (so to speak), and that the affordances of using thin-slice methods can significantly lower the time and burden of reading and scoring. In sum, we concluded that thin-slice methods can be added to the Writing Studies toolkit for large-scale direct assessment of evaluative reflective writing.

Qualitative Findings

To move to the qualitative component of our study, we first looked at the frequency distribution scores in the rubric categories (see Table 2).

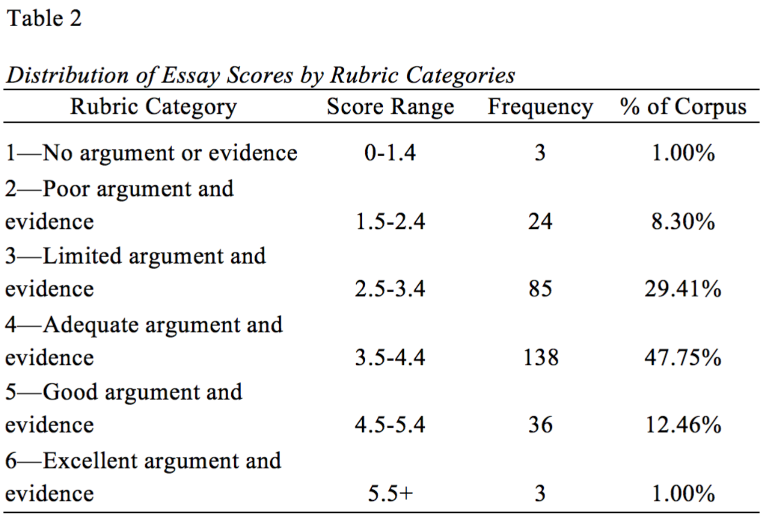

Table 2

Distribution of Essay Scores by Rubric Categories

Rubric Category	Score Range	Frequency	% of Corpus
1—No argument or evidence	0-1.4	3	1.00%
2—Poor argument and evidence	1.5-2.4	24	8.30%
3—Limited argument and evidence	2.5-3.4	85	29.41%
4—Adequate argument and evidence	3.5-4.4	138	47.75%
5—Good argument and evidence	4.5-5.4	36	12.46%
6—Excellent argument and evidence	5.5+	3	1.00%

Table 2 *Distribution of Essay Scores by Rubric Categories*

We found essays scored as Adequate or Good featured well-developed arguments focused on the course learning outcomes and course concepts; these student writers also wrote reflectively about changes over time and offered supporting evidence for their reflective claims. For example, in essays that were scored Adequate or Good, students named and discussed course outcomes and concepts, provided details about their progress made in pursuit of the outcome, described their learning process, and/or provided quotations from their own writing as evidence of learning. Conversely, essays that were scored in the Poor or Limited categories spent little to no time discussing learning outcomes or course concepts; also, these student writers did not write reflectively about changes in their writing over time (i.e., they did not reflect about their struggle, improvement, or progress in the course), nor did they provide sufficient evidence for their statements and claims.

In sum, student work that scored in the Poor and Limited categories can be contrasted to work in the Adequate and Good categories on the weakness or strength of the argument and evidence. For example, thesis-driven essays in the genre of academic argument were rated higher than essays offering narratives or personal responses. Additionally, paragraph development and evidence-based development differed across the essays; those that were scored Adequate or Good typically had well-developed body paragraphs that included specific evidence to support claims about achieving course outcomes whereas essays that were scored Limited or Poor typically provided only vague generalizations concerning the writing process and made only loose connections between their actions and achievement of the learning outcome.

If one goal of assessment is to “move the needle” so more students achieve the Reflection outcome in our FYC course, the qualitative analysis indicated that low-ranked reflective essays often neglected basic elements of argumentation, frequently failed to make concrete and significant claims based on looking back at one's own learning and writing experiences, and made no explicit connection between claims and evidence of reflective learning. They also did not address how reflection can or did function as a process that leads students to regulate thinking or writing. There was also little to no evidence in these low-scoring essays of students’ abilities to connect reflection to the critical thinking that is so important to the university experience.

Discussion

We concluded above that thin-slice methods can reliably be used as a quantitative method in the large-scale direct assessment of evaluative reflective essays. Here, we return to the issues of validity raised in the introduction: Is there any evidence indicating that the methods and findings of this study were valid, and, if so, what kind(s) of related validity have been achieved? To contextualize this question, we must first emphasize that the thin-slice methods used and tested here are specific to our particular context—an exigence of sustainability and a set of methodological concerns. Our previous direct assessments of student writing were not based upon a representative sample or reliable scoring, so we were making programmatic decisions about our FYC curriculum without data-driven support from our assessment practices.

We are emphatically not making a claim that thin-slice assessment methods are right for every assessment context. We are, however, hoping to make a strong argument that the use of mixed methods, including thin-slice methods, offers important affordances to the field of Writing Studies. In our view, too many direct assessment studies are not based upon representative samples, nor are they based upon mixed methods approaches when appropriate, a perspective shared by Haswell (2012) and by White et al. (2015):

> we need to be clear about our reason for advocating empirical techniques of program assessment: Preventing over-simplified measures of writing programs—measures that fail to support instruction—is simply not possible without the use of sophisticated quantitative and qualitative processes … The use of descriptive and inferential statistics is more effective in administrative meetings than the rhetoric often employed. (White et al., 2015, p. 114)

If we are to use quantitative methods, though, we must come to terms with reliability and validity within this domain of formal assessment.

Institutional Implications and Implementation

Based on our findings, we developed a series of recommendations for curricular reform, which we hope will result in improved instruction, professional development, and student achievement in our FYC course. First, we recommended making clearer distinctions between narrative and argument genres of reflective writing and placing greater emphasis and instructional scaffolding around the genre of evaluative argumentation in reflective writing. In other words, when we ask students to reflect in this manner, we need to be more explicit about advising students to make strong claims about changes they have made throughout their course experience or changes they made in relation to a particular outcome for the course. We also need to give more attention to the relationship between reflection and argument so that reflective writing is partly conceived through a rhetorical framework that encourages the use of reflection in “learning how to learn” in a composition course. More specifically, our team recommended that reflective writing be integrated to include short post-project reflective assignments prompting students to reflectively practice stasis genres (e.g., arguments based on making evaluations and developing definitions or identifying cause/consequence relationships). We further suggested that instructors provide written feedback on this series of short reflective writing activities.

Second, our rich-features analysis identified a related implication about supporting the reflective essay assignment with more explicit attention to teaching paragraph-level expressions of students’ ideas about the role of reflection in their course. We thus recommended that instructors emphasize basic paragraph development, focusing on unified expression of specific claims and adequate evidence related to course concepts, classroom discussions, or other analyses, examples, or experiences that advanced their understanding of college writing. Emphasizing paragraphing in the reflective essay can also reinforce how different forms of reflective knowledge can support the larger argument. For example, students might draft paragraphs that demonstrate declarative knowledge (about concepts, facts, skills, or subject matter that can impact student performance), procedural knowledge (about how heuristics or elements of the writing process can impact student performance) or conditional knowledge (about how and when to apply course concepts or procedures to improve student performance).

Third, it became known to the Research Team that there was considerable variety in both the amount of time instructors spend introducing the reflective essay assignment and the instructional strategies used to support it. We thus recommended that we work toward greater instructional uniformity across sections of the course, particularly in terms of how much instructional time is dedicated to the reflective essay assignment in the course schedule. We also recommended that we review the assignment prompt for the reflective essay and develop an assignment-specific grading rubric for reflective essays.

The assessment findings and our recommendations in this study were summarized in a memo and forwarded to our program’s Curriculum Committee, the body responsible for curricular reform in our program. In this way, we did not privilege the Research Team’s views over the shared governance of the program as represented in our committee structures and processes, thereby empowering the internal stakeholders of our program—the instructors of our FYC course—to use assessment information to design and implement curricular reform aimed at positively impacting our students’ reflective writing abilities as would be demonstrated in their end-of-semester reflective essays.

Limitations

There were several limitations to our study. First, we examined only one outcome (Reflection) in our assessment process. In future research, our program plans to continue testing and refining thin-slice assessment methods for written language in our yearly assessments. Such studies would aim to test thin-slice methods when assessing for other outcomes such as reading or research. Similarly, our study focused on the ways that the use of a thin-slice technique can reduce time spent on assessment in large writing programs. Given this focus, we did not examine whether the method might also produce gains beyond the increase in efficiency and IRR that resulted in that context. It should, however, be kept in mind that representative sample sizes do not grow proportionally with the size of the set from which the sample is being extracted. In other words, since the percentage of texts needed for explicit assessment of student writing is actually lower when being extracted from a larger total set, assessing a representative sample can create a disproportionately larger burden for mid-size and smaller writing programs. For those reasons, thin-slice techniques are likely to increase efficiency for programs of varying sizes as long as they are using a representative sample as part of their assessment process. That said, we did not test specifically for efficiencies or improvements that might be equally advantageous to assessment in relation to smaller sample sizes. One avenue of investigation that might be pursued in that context would be studying the impact of thin-slice approaches on ease of scoring. Based on research indicating that length of a writing sample negatively impacts ease of scoring in traditional approaches to writing assessment (Wolfe, Song, & Jioa, 2016), one might hypothesize that the thin-slice technique might improve the IRR of a rating team of any size because it decreases and standardizes the samples being assessed. Having not investigated those specific conditions as part of this study, however, we can only at this time presume this effect.

Another limitation to this study was the discovery of a need to develop assessment materials and procedures such as rubrics and norming designed purposefully for thin-slice assessment of written language. This apparent limitation became clear late in this research project as plans for further testing of thin-slice methods in writing assessment were discussed. As research extends and adapts thin-slice assessment methods into written language, an infrastructure of assessment (rubrics, norming protocols, etc.) sensitive to new methods will emerge. For example, we are now developing a study to test specialized norming for the thin-slice assessment of reflective essays.

Implications for the Field

Beyond the expected contributions of this study to curricular changes and student performance in our own program, we believe this project also makes a significant contribution to existing scholarship and identifies best practices in the large-scale direct assessment of reflective writing in FYC. Through our mixed methods approach to assessment, we were able to demonstrate a feasible method for achieving not only reliability and validity but also validation, via the high level of IRR and consistency in the quantitative analysis and in the consequential and curricular-focused process of our qualitative analysis (White et al., 2015). In other words, we maintain thin-slice methods can contribute a valuable base of quantitative evidence, including statistical reliability and validity, in order to pursue the higher-level validation sought by assessment researchers in Writing Studies.

Even more notably, we were able to achieve these goals via a process that was highly efficient in its required time and resources. Indeed, while the time spent per rater in this assessment study was significantly lower than in our typical assessment process (specifically, the Research Team's average assessment time per rater was less than half the time per rater of the Regular Team), thin-slice scores had a positive correlation with the Regular Team scores and an even higher degree of inter-rater reliability than the Regular Team. This efficiency in turn makes it feasible for WSU's Composition Program to assess a fully representative sample of student writing, something it had failed to achieve due to resource limits in the past. Finally, in successfully piloting thin-slice methods in writing assessment, we have offered the field a new and potentially very useful assessment method for composition and writing programs with large student enrollment. Taken together, these processes open pathways for sustainable assessment methods that might allow us to achieve the "best of both worlds" in regard to contemporary debates over programmatic autonomy and programmatic accountability, the value of quantitative and qualitative assessment methods, and the evidence of validity within the framework of validation.

References

Ambady, N., Bernieri, F. J., & Richeson, J.A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin-slices of the behavioral stream. In M. Zannna (Ed.), *Advances in experimental social psychology* (Vol. 32, pp. 201-272). New York, NY: Academic Press.

Ambady, N., Koo, J., Rosenthal, R., & Winograd, C. H. (2002). Physical therapists' nonverbal communication predicts geriatric patients' health outcomes. *Psychology and Aging, 17*(3), 443-452. doi:10.1037/0882-7974.17.3.443

Ambady, N., Krabbenhoft, M. A., & Hogan, D. (2006). The 30-sec sale: Using thin-slice judgments to evaluate sales effectiveness. *Journal of Consumer Psychology, 16*(1), 4-13. doi:10.1207/s15327663jcp1601_2

Ambady, N., LaPlante, D., Nguyen, T., Rosenthal, R., Chaumerton N., & Levinson, W. (2002). Surgeons' tone of voice: A clue to malpractice history. *Surgery, 13*(2), 5-9.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin-slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*(3), 431-441. doi:10.1037/0022-3514.64.3.431

Barton, E. (2004). Linguistic discourse analysis: How the language in texts works. In C. Bazerman & P. Prior (Eds.), *What writing does and how it does it* (pp. 57-82). New York, NY: Routledge.

Behizadeh, N. (2014). Mitigating the dangers of a single story: Creating large-scale writing assessments aligned with sociocultural theory. *Educational Researcher*, *43*(3), 125-136. doi:10.3102/0013189x14529604

Carcone, A., Naar, S., Eggly, S., Foster, T., Albrecht, T., & Brogan, K. (2015). Comparing thin slices of verbal communication behavior of varying number and duration. *Patient Education and Counseling, 98*(2), 150-155.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*(3), 205-215.

Cherry, R. D., & Meyer, P. R. (1993). Reliability issues in holistic assessment. In M. M. Williamson & B. A. Huot (Eds.), Validating holistic scoring in writing assessment: Theoretical and empirical foundations (pp. 109-141). Cresskill, NJ: Hampton.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284-290. doi:10.1037/1040-3590.6.4.284

College Board. (n.d.). *Validity evidence: Types of validity evidence*. Retrieved from https://research.collegeboard.org/services/aces/validity/handbook/evidence

Condon, W. (2011). Reinventing writing assessment: How the conversation is shifting. *Writing Program Administration, 34*(2), 162-182.

Elbow, P. (2012). Good enough evaluation: When is it feasible and when is evaluation not worth having? In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Ed White* (pp. 301-323). New York, NY: Hampton Press. Retrieved from https://works.bepress.com/peter_elbow/46/

Elliot, N. (2015). Validation: The pursuit. *College Composition and Communication*, *66*(4), 668-685.

Elliot, N., Plata, M., & Zelhart, P. (1990). *A program development handbook for the holistic assessment of writin*g. Lanham, MD: University Press of America.

Gladwell, M. (2005). *Blink: The power of thinking without thinking*. New York, NY: Little, Brown.

Gorzelsky, G., Driscoll, D., Pazcek, J., Hayes, C., & Jones, E. (2016). Metacognitive moves in learning to write: Results from the Writing Transfer Project. In J. Moore & C. Anson (Eds.), *Critical transitions: Writing and the question of transfer*. Retrieved from https://wac.colostate.edu/books/ansonmoore/

Grahe, J. E., & Bernieri, F. J. (1999). The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior, 23*(4), 253.

Haefner, J. (2011). *A practical, pithy guide to quantitative scoring assessment at a SLAC*. Retrieved from https://sun.iwu.edu/~jhaefner/practical_pithy_guide.pdf

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-24.

Haswell, R. H. (2012). Fighting number with number. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Ed White* (pp. 413-423). New York: Hampton Press.

Henry, S. G., & Eggly, S. (2013). The effect of discussing pain on patient-physician communication in a low-income, black, primary care patient population. *The Journal of Pain, 147*, 759-766.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Houghton Mifflin.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, *47*(4), 549-66.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research, 2001*(109), 9-25. doi:10.1002/ir.1

Lippa, R. A., & Dietz, J. K. (2000). The relation of gender, personality, and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior, 24*(1), 25.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: American Council on Education [ACE]/Macmillan.

Moss, P. (1992). Shifting conceptions of validity in educational measurement. *Review of Educational Research*, *62*(3), 229-258.

Moss, P. (1994). Can there be validity without reliability? *Educational Research, 23*(4), 5-12.

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69-71.

Murphy, N. A. (2005). Using thin-slices for behavioral coding. *Journal of Nonverbal Behavior, 29*(4), 235-246. doi:10.1007/s10919-005-7722-x

Murphy, N. A., Hall, J. A., & Colvin, C. R. (2003). Accurate intelligence assessments in social interactions: Mediators and gender effects. *Journal of Personality, 71*(3), 465-493.

National Center for Educational Statistics. (2016). *Integrated postsecondary education data system (IPEDS) data center* [Data set]. Retrieved from https://nces.ed.gov/ipeds/datacenter/

National Statistical Service. (n.d.). *Sample size calculator*. Retrieved from http://www.nss.gov.au/nss/home.nsf/pages/Sample+size+calculator

Office of Budget, Planning, and Analysis. (n.d.). *Wayne State University peer list*. Retrieved from https://budget.wayne.edu/institutional-research/peer-list

O’Neill, P., Moore, C., & Huot, B. (2009). *A guide to college writing assessment*. Logan, UT: Utah State University Press.

Peracchio, L. A., & Luna, D. (2006). The role of thin-slice judgments in consumer psychology. *Journal of Consumer Psychology, 16*(1), 25-32. doi:10.1207/s15327663jcp1601_5

Redwine, T., Leggette, H. R., & Prather, B. (2017). A case study of using metacognitive reflections to enhance writing skills and strategies in an agricultural media writing course. *Journal of Applied Communications, 101*(1), 56-68. doi:10.4148/1051-0834.1014

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

Taczak, K. (2015). Reflection is critical for writers’ development. In L. Adler-Kassner & E. Wardle (Eds.), *Naming what we know: Threshold concepts of writing studies* (pp.75-76). Boulder, CO: The University Press of Colorado.

Tinberg, H. (2015). Metacognition is not cognition. In L. Adler-Kassner & E. Wardle (Eds.), *Naming what we know: Threshold concepts of Writing Studies* (pp.75-76). Boulder, CO: The University Press of Colorado.

Vitiello, G. (2016, February 22). What should you do during the year to stay class reliable? [Web log post]. Retrieved from http://info.teachstone.com/blog/stay-class-reliable-through-the-year

White, E. (2005). The scoring of writing portfolios: Phase 2. *College Composition and Communication, 56*(4), 581-600.

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

Wolfe, E. W., Song, T., & Jioa, H. (2016). Features of difficult-to-score essays. *Assessing Writing*, *27*, 1-10.

Yancey, K. B. (1998). *Reflection in the writing classroom*. Logan, UT: Utah State University Press.

Appendix A

Reflective Essay Assignment Description

Introduction/Rationale. The Reflective Essay is a 1000-1250-word (4-5 pages) essay in which you make a case for how well you have achieved the goals of the course. To do so, you must look back over the work you produced during the semester in order to find, cite, and discuss evidence of achievement in each of the four learning outcome categories (reading, writing, research, and reflection. It is critical that your Reflective Essay includes concrete examples and discussion of what you have been able to do as a result of your work in the course.

While your discussion of achievements with respect to ENG 1020 learning outcomes is perhaps the most important goal in the Reflective Essay, the written expression of these achievements can be strengthened when it is integrated into a broader narrative that describes where you are coming from and who you are as a student. In this narrative, you may discuss, for example, how you learned and used various reading strategies in the course, or you may describe, for example, how your ability to perform effective research increased.

In sum, the Reflective Essay should make claims about your success with respect to ENG 1020 learning outcomes and support these claims with compelling evidence of achievement in order to demonstrate what you have learned and what you can do as a result of your work in the course. In this way, a successful Reflective Essay will inspire confidence that you are prepared to move forward into your next composition courses and into the larger academic discourse community.

Assignment Prompt: In this assignment, you will evaluate your growth as an English 1020 student, using your choice of experiences and work on the projects to support your claims. In an essay of 4-5 pages, make an argument that analyzes your work in ENG 1020 in relationship to the course learning outcomes listed on the syllabus for the course. Explain what you have achieved for the learning outcomes by citing specific passages from your essays and other assigned writings for the course, and by explaining how those passages demonstrate the outcomes. Also, consider describing the process you used to complete this work and any background information about yourself, as listed above, that might help us better understand the work you did this semester in working toward the course learning outcomes.

You will want to choose the learning outcomes and knowledge that have developed most strongly and importantly for you. If you think there is little evidence of your growth in a particular learning outcome, no problem: just articulate why in your final essay. You should address all of the learning outcomes, but you may choose which ones you focus on. Your main claim (or thesis statement) should identify specific characteristics that you believe your experiences and work in English 1020 (which you'll use your body paragraphs to talk about) represent. The body paragraphs of your essay should develop your main claim with evidence from your major works and experiences in this course. As you choose evidence and sub-claims to make about your major assignments, you will develop your paragraphs by drawing upon the process of completing the assignment to support the claim.

In a nutshell, this assignment asks you to take a critical look at your work from this semester, and talk about it in terms your knowledge of yourself as a learner and thinker.

Appendix B
Reflective Essay Rubric

	6	5	4	3	2	1
Learning Outcome		**6, 5, 4 Sufficient**			**3, 2, 1 Insufficient**	
Use written reflection to evaluate one's own learning and writing.	**Excellent** argument (thesis, claim, relation to course outcomes). **Excellent** evidence (examples, analysis, experiences, discussion).	**Good** argument (thesis, claim, relation to course outcomes). **Good** evidence (examples, analysis, experiences, discussion).	**Adequate** argument (thesis, claim, relation to course outcomes). **Adequate** evidence (examples, analysis, experiences, discussion).	**Limited** argument (thesis, claim, relation to course outcomes). **Limited** evidence (examples, analysis, experiences, discussion).	**Poor** argument (thesis, claim, relation to course outcomes). **Poor** evidence (examples, analysis, experiences, discussion).	**No** argument (thesis, claim, relation to course outcomes). **No** evidence (examples, analysis, experiences, discussion).

Logical Decision Rule: if an essay has a mixed score, record the lower score.
Example: if an essay has excellent evidence and a good argument, then it is a 5/Good.
Example: if an essay has an adequate argument but poor evidence, then it is a 2/Poor.

Biosketches

All authors were members of the 2015-2016 and 2016-2017 **Composition Research Committee** in the English Department of Wayne State University, Detroit, Michigan.

Jeff Pruchnic is Associate Professor and Director of Composition in the Department of English at Wayne State University. His work on rhetorical theory and writing pedagogy has appeared in *JAC*, *Rhetoric Review*, *Rhetoric Society Quarterly* and elsewhere.

Chris Susak is a Lecturer and PhD candidate in the Department of English at Wayne State University. His research interests include community writing, writing pedagogy, and rhetorical theory.

Jared Grogan is a Senior Lecturer in the Department of English at Wayne State University. His research interests include the history of rhetoric, science studies, ecology, technical communication, computers and composition, and writing pedagogy.

Sarah Primeau is a PhD student in the Department of English at Wayne State University. Her research interests include cultural rhetorics, first year writing pedagogy, and assessment.

Joe Torok is a Lecturer and PhD student in the Department of English at Wayne State University. His research interests include cultural rhetorics, visual rhetorics, and network studies.

Thomas Trimble is a Senior Lecturer in the Department of English at Wayne State University. His interests include writing assessment, community-based learning, and first year writing pedagogy.

Tanina Foster is the Assistant Director of the Behavioral and Field Research Core in the Population Studies and Disparities Research Program at Karmanos Cancer Institute/Wayne State University. Her research interests include understanding current and new applications of research methodology and techniques and their specific application to the understanding of human behavior, communication patterns and the subsequent impact on health care decision-making.

Ellen Barton is Professor of Linguistics and English and Associate Provost and Associate Vice President for Academic Personnel at Wayne State University. Her research interests include mixed methods research methodologies, and she received the 2009 *College*

Composition and Communication Richard Braddock Award for her article “Further Contributions from the Ethical Turn in Composition/ Rhetoric: Analyzing Ethics in Interaction.”