Journal of Writing Assessment

Prebel, J. & Li, J. (2024). Multifaceted equity: Critiquing a first-year writing assessment through curricular, performance, and reliability lenses. *Journal of Writing Assessment*, *17*(2). **DOI**: 10.5070/W4.jwa.1555

Multifaceted Equity: Critiquing a First-Year Writing Assessment Through Curricular, Performance, and Reliability Lenses

Julie Prebel, Occidental College, US, jprebel@oxy.edu
Justin Li, Occidental College, US, justinnhli@oxy.edu

Abstract: This article examines whether a college's new portfolio-based first-year writing assessment process is fair, equitable, and antiracist. We build on the existing literature by arguing that equity must be assessed through multiple facets within the assessment ecology and operationalize three lenses through which to examine the assessment process. The curricular lens shows that the new writing assessment is more aligned with and improved the classroom pedagogy of our first-year seminars. The performance lens shows the ongoing disparities between students across demographic backgrounds. Finally, the reliability lens reveals faculty differences in how they interpret the assessment criteria. We conclude that while the new portfolio assessment is fairer and more equitable for most students, it is also constrained by institutional structures and process elements that prevent it from being antiracist.

Keywords: equity, first-year writing, assessment, portfolio, student demographics

Introduction: A Move Towards Equitable and Antiracist Writing Program Assessment

This article emerges from and contributes to scholarly conversations about equitable, socially just writing assessment practices in two central ways. First, we deepen the conversation on equitable writing assessment by claiming that equity contains multiple facets and that an evaluation of an assessment practice for equity must examine more than one of these. We look at three facets specifically: how the assessment integrates with the curriculum (the curricular lens), how different student populations perform on the assessment (the performance lens), and how consistent raters are in assessing writing (the reliability lens). Second, we present a case study applying these three facets to a revised writing proficiency assessment at our college and discuss how changing our process did and did not result in a more equitable assessment ecology.

We draw the key terms we use to describe our aims, practices, and results—fairness, equity, and antiracist—from composition scholarship on social justice and writing assessment. Scholars have written about the challenges of creating and sustaining writing assessments that align with institutional requirements while being fair practices that do not disadvantage students. Fairness has become a unifying concept in writing assessment, as Poe and Inoue (2016) note, providing a way for composition studies to consolidate our various empirical research into a broader conversation about enacting equitable initiatives. Yet, as we discuss in the analysis of our study, fairness and equity, while connected, are not necessarily synonymous; assessment practices can be fairer by implementing clearer standards or unifying criteria, but there may be assessment situations when fairness prevails without equitable outcomes. Equity, too, might be understood as an umbrella term that refers directly or obliquely to the effects of sociocultural variables such as gender, race, and linguistic standards on assessment practices. Antiracism may be part of the broader term "equity" and enacting an antiracist assessment would thus be one way to achieve an equitable assessment, but an equitable assessment may not be antiracist.

In changing our assessment, we were motivated to implement a process that would result in fairer outcomes for students, which we hoped would not reify racial formations resulting in inequities (Anson, 2012; Hammond, 2019; Inoue, 2015; Inoue & Poe, 2012; Prendergast, 1998; White, 1994). We paid particular attention to Anson's (2012) claim that writing across the curriculum (WAC) assessment processes have neglected diverse learners more broadly and race and racial identities more specifically and Inoue's (2015) reminder that even in discussions of justice and fairness there is a tendency to "avoid the racial" (p. 5). Increasingly, scholars have amplified the recognition of racial hierarchies and processes of racialization that have an impact on writing program administration and assessment (Branson & Sanchez, 2021; Carter-Tod, 2019; Carter-Tod & Sano-Franchini, 2021; Gere et al., 2021; Perryman-Clark, 2016) and argued for understanding assessment practices and outcomes as shaped by and reflective of racial constructs and experiences (Inoue & Poe, 2012, p. 6).

Designing a fair and equitable writing assessment practice thus prompted us to rethink an ecology in which racism and other inequalities inevitably existed. In examining the racial *habitus*, or the function of race in writing assessment, Inoue (2015) underscores the ways we hold onto myths about the ideal student, their ideal writing performance, and, consequently, our idealization of assessment outcomes. The myths and models of what constitutes "good" writing create advantages for students who conform to the "white racial set of experiences and perspectives" while disadvantaging

students who may not use the "dominant discourses" that tend to be "historically connected" to a white racial *habitus* (Inoue, 2015, pp. 28–30). Developing antiracist writing assessment requires investigating our practices for evidence of racism and racialized power dynamics that are already present in our social and institutional structures (Inoue, 2015). In our analysis below, we address ways that our process—both before and after the assessment changes—is enmeshed in racialization and other biases that challenged our goal of creating an equitable assessment that is also antiracist.

There is considerable literature on equitable practices in writing assessment grounded in student, faculty, and institutional contexts (Hamp-Lyons & Condon, 1993/2009; Inoue & Poe, 2012; Inoue, 2015; Poe et al., 2018; Stewart, 2022), but we need more examples evaluating *in situ* assessment that contextualize assessment results from multiple frames of reference. Through our three lenses—curricular, performance, and reliability—we analyze the entirety of our "assessment technology" (Inoue & Poe, 2012, p. 3), or the interconnected elements that form our writing assessment. By approaching our investigation of our assessment in this way, we provide a comprehensive or holistic view of the process as we consider how to assess writing and design assessment to address issues of fairness and equity. As we show below, we have moved closer to realizing a vision of an equitable writing assessment practice, but our process also raises questions about *what it means* for writing assessment to be equitable. Moreover, our study proves valuable to the field by providing an example of holistic analysis of an assessment process, including disaggregated data analyses, that others might apply in assessing their programs.

This article presents a case study evaluating the changes we made to the writing proficiency assessment for all first-year students at Occidental College, a small liberal arts college. Concerns with our first-year assessment, as explained below, led to changes beginning in the academic year 2019-2020. In evaluating this new process, we collected data on our assessment results for four years before the change (since AY 2015-2016) and for four years after the change (through AY 2022-2023).

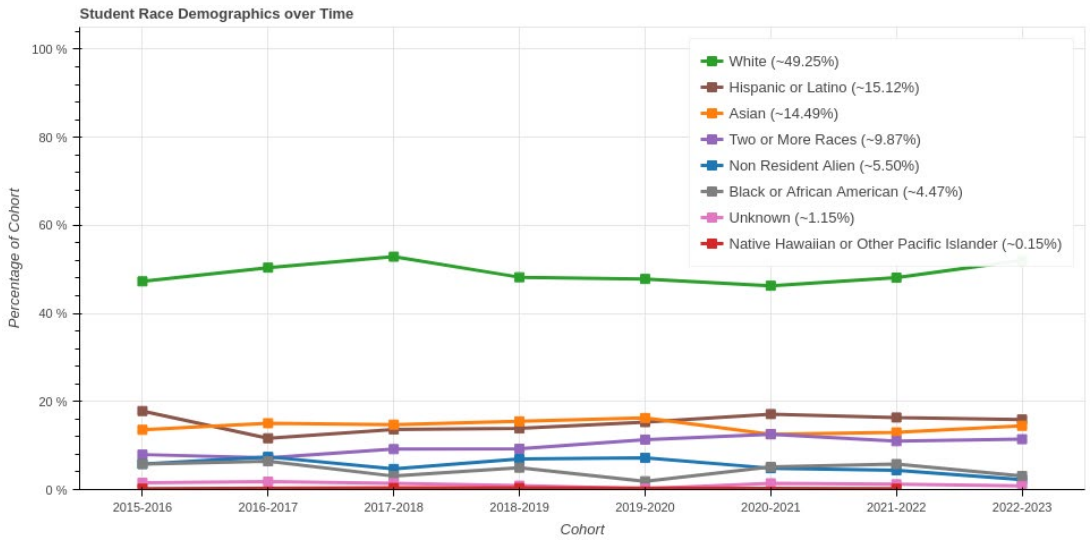
Institutional Contexts: From Timed Writing Exams to Portfolio Assessment

At our college, all students must complete a two-stage writing proficiency requirement to graduate. The first-stage requirement introduces students to college-level writing expectations with a focus on expository essays and prepares students for more complex writing tasks within an academic discipline. All students take two First Year Seminar (FYS) courses over two semesters, which are taught by faculty from a range of disciplines, consistent with a WAC model. The number of FYS courses tracks with the admitted students for a given year (~520 per), and typically, we have 60-65 sections taught by roughly the same number of faculty (with some teaching more than one section). In the eight year period we examine, the demographics of our student population taking FYS have remained consistent, with a higher number of female (58.77%) to male students (41.23%) averaged across all years; we note that the college started reporting class profile information by female, male, and non-binary only recently and not in the eight-year period we discuss. Of all students in our analysis period, 13.49% are first-generation; see Figure 1 for racial demographics of students in this same analysis period of AY 2015-2016 through AY 2022-2023.

Prior to AY 2019-2020, to complete the first-stage requirement, students received three assessment scores that indicated proficiency through a six-point holistic rubric (along with passing both semesters of FYS). Students received an assessment score from their fall FYS instructor, their spring FYS instructor, and through a 55-minute timed writing exam (TWE) taken in the spring of

Figure 1

Racial Demographics of Occidental College, AY 2015-2016 to AY 2022-2023



their first year and scored by a committee of faculty assessors. If students scored a 4 or higher on the six-point rubric in two of these three measures, their writing was deemed proficient and they completed the requirement. Students scoring a 3 or below on two or more measures were required to take a 200-level writing course in their second year.

A review of this assessment process raised questions about our measurement standards and the lack of alignment of our assessment with scholarship and best practices in composition studies. We identified concerns with key elements of our assessment ecology:

- Some FYS courses offered minimal writing instruction, resulting in unpredictable and subjective faculty assessment scores.
- There was a lack of correlation between FYS assessment scores and TWE assessment scores.
- The use of the TWE to determine proficiency was outdated in composition studies and did not align with the writing students were asked to produce in FYS.
- Our assessment process did not provide opportunities for student reflection on their writing knowledge and process.

In changing the assessment to address these concerns, we focused especially on the problems with timed essay examinations, well-documented in composition scholarship (Elbow & Belanoff, 1986/2009; Huot, 1996; Moss, 1994/2009; Lau, 2013; Petersen, 2009). Knowing that writing proficiency exams are neither accurate nor valid as an assessment of a student's writing ability, we sought to create a new process denotative of Huot's (1996) five principles that an assessment be site-based, locally-controlled, context-sensitive, rhetorically-based, and with procedures that are accessible to all stakeholders (see also CCCC, 2022).

We aligned our new assessment with these best practices to include writing tasks linked directly to the writing goals and curriculum of our FYS program. In the new assessment process, students must still pass their FYS courses. However, we replaced the FYS faculty assessment

scores and the TWE with a portfolio of three essays students select from their FYS courses and a Reflective Introduction Essay, common elements in portfolio-based writing systems (Elbow & Belanoff, 1986/2009). We replaced the six-point holistic rubric with a five-criteria portfolio rubric:

1. Reflection: students' self-reflection of their writing processes and products.
2. Focus: assignment responsiveness and thesis clarity.
3. Organization: cohesion and paragraph structure.
4. Evidence & Development: source use and idea development.
5. Writing Features & Presentation: prose, grammar, and mechanics.

Portfolios are read by two faculty accessors, with a third to adjudicate any disagreements. Students can revise their essays with support from the college's Writing Center before submitting their portfolio, giving them more agency and encouraging them to produce "better writing" (Elbow & Belanoff, 1986/2009, p. 98), and 42-50% of students have made use of the Writing Center in revising their portfolio essays.

In the sections below, we provide data on the interconnected elements of our assessment ecology through curricular, performance, and reliability lenses, and enact a conscientious analysis of our new practice to consider whether it is fair and equitable (Inoue, 2015). We provide a holistic analysis that allows us to identify specific places where the changes we made resulted in fairer and in some cases more equitable results, and where elements in our process were limitations to enacting an antiracist assessment—and thus a more comprehensively equitable one.

Curricular Lens

In this section, we consider the relationship between our FYS curriculum and students' writing assessment results. In making changes to our process and evaluating those changes, we are guided by scholarship confirming that a portfolio-based assessment improves not only the assessment process itself but also faculty pedagogy and curriculum development (Hamp-Lyons & Condon, 1993/2009; Weiser, 1997; Yancey & Weiser, 1997). We examine how the shift from the TWE to a portfolio required changes to our FYS program and the connection between classroom writing experiences and fairer assessment outcomes. Using a curricular lens, we consider how an equitable writing assessment practice involves developing a "common interest in student achievement" (Anson, 2012, pp. 26–27), which we have tried to reinforce through revised curricular elements aimed to improve student writing outcomes. While a detailed evaluation of the specific FYS faculty pedagogies is beyond the scope of this article, we look at how the changes to our practice created more alignment between classroom teaching and evaluation and programmatic assessment. We consider, too, the considerable scholarship on how changes to classroom writing assessment practices can engender antiracist assessment projects both in and outside of the classroom (Inoue, 2015, p. 10). For example, in applying antiracist theory and pedagogy to a study of the writing classroom, Inoue (2015) offers ways to think about how classroom ecologies influence and offer possibilities for large-scale writing assessment (p. 12). We consider these synergies between classroom and programmatic assessment by looking at how our revised college-wide assessment process necessitated changes to the FYS curricular requirements. The data we analyze suggest that the intersecting curricular and programmatic elements of our new process encourage fairness and equity and may move us a step closer to enacting an antiracist assessment.

In changing our assessment process, we looked carefully at our FYS program and identified elements that created a disconnect with our assessment outcomes, especially variability in the

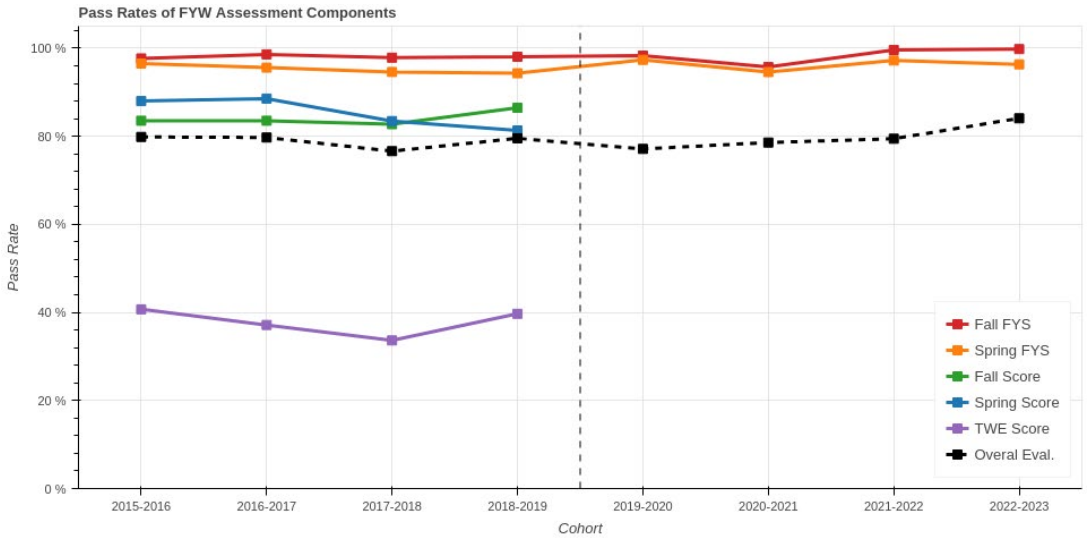
curriculum across the FYS courses, subjective faculty assessments of students' writing proficiency, and lack of alignment of faculty pedagogy with the assessment process. While the overall WAC approach to staffing our FYS courses has not changed, one main concern we noted of our previous assessment practice was that there were few curricular requirements of the faculty teaching the courses; mainly, faculty were asked to assign a minimum number of writing assignments (four) and to provide (at least some) instruction to help students develop writing skills and knowledge. We found that these broad curricular suggestions were applied unevenly, with some FYS sections focused on writing while others prioritized non-writing content, resulting in varying student learning of writing concepts and transferable writing knowledge. Moreover, classroom writing instruction did little to prepare students for the TWE, as it was challenging for faculty (and not that appealing) to include a curricular element to teach students to write in a timed environment. Our TWE, similar to the concerns widely noted in the scholarship on proficiency examinations, also offered no opportunity to write more than one draft, to revise work in response to feedback, or to reflect on the writing process (Elbow & Belanoff, 1986/2009). In sum, we found that the TWE sent the "wrong message about the writing process" and what constituted proficient writing (Elbow & Belanoff, 1986/2009, p. 97; see also Trimbur, 1996; Yancey & Weiser, 1997). We also found a lack of correlation between faculty evaluation of students' writing—course grades and rubric assessments—and students' TWE writing scores.

Changing the assessment to remove the element of faculty rubric scores and replacing the TWE with a writing portfolio resolved many of these alignment issues. Figure 2 quantifies the misalignment between TWE, the FYS holistic scores, and the FYS grades.

As Figure 2 shows, before implementing the portfolio in AY 2019-2020, there is disconnection between fall and spring scores and the TWE. Whereas the FYS instructor scores show high pass

Figure 2

Pass Rates for Each Component of the First-Year Writing Requirement



Note. Prior to AY 2019-2020, the writing assessment consists of three components, two of three of which must be passing for completion of the requirement. After AY-2019-2020, the assessment consists only of the portfolio, and thus no separate line is shown.

rates around 85% (the green and blue lines), the majority of students failed the TWE, with a pass rate of only 38-40% (the purple line). Since only two passing scores from fall, spring, and the TWE were required to complete the First-Stage writing requirement, the overall pass rate is around 80% (the black dashed line). This discrepancy highlights the misalignment of the TWE. Starting in AY 2019-2020, the portfolio became the only component of the assessment, and therefore its pass rate is the same as that of the overall First-Stage writing requirement (the black dashed line). We were not aiming for a specific pass rate, but the portfolio maintained a pass rate of around 80%, the same as before. While pass rates alone do not tell the full story, this data suggests that the portfolio resolved the misalignment of the TWE while continuing to capture the students' aggregate writing proficiency.

Switching to the portfolio also resulted in positive changes to the FYS curriculum. We aligned the FYS writing outcomes with the portfolio criteria, which helped mitigate the disconnect between the standards used to evaluate students' classroom writing and the assessment of their portfolio essays, a concern commonly noted in portfolio-based assessment (Hamp-Lyons & Condon, 1993/2009). We also developed clearer writing expectations for the FYS courses and asked instructors to design curricula to prepare students for portfolio requirements, including assigning at least one thesis-driven essay and teaching students to write a reflective essay—components that create greater connection between classroom writing experiences and portfolio expectations. Adapting to a portfolio system and to concurrent curricular expectations initially made instructors uncertain about how to implement the new requirements—another concern noted in the literature on portfolio assessment (Elbow & Belanoff, 1986/2009)—which we addressed through enhancements to our faculty development workshops, with more emphasis on teaching the writing elements assessed in the portfolio rubric, reflective writing, and orienting specific faculty or course goals with program-wide learning outcomes.

In addition to the curricular benefits of professional development through these workshops, teaching practices have improved through training faculty to be portfolio assessors. Prior to the portfolio shift, typically fewer than 50% of FYS faculty participated in the evaluation of the timed writing exams. Just as there was a disconnect between the writing students produced in class and their TWE essay, there was little connection between the TWE evaluation process and faculty writing pedagogy. Beginning in AY 2019-2020, the majority of FYS faculty now serve on the Portfolio Assessment Committee and participate in at least three professional development workshops that prepare them to assess and reflect on the assessment of student portfolios. In these workshops, faculty gain knowledge about the portfolio requirement, discuss their responses to sample portfolios in a norming session, and debrief after completing the assessment. As faculty learn about the assessment process and familiarize themselves with the portfolio rubric criteria, they acquire or enhance their understanding of writing concepts they can use in their pedagogy and classroom assessment practices, as noted in assessor feedback:

> I used the portfolio rubric exclusively for providing feedback on essays in my FYS class. I gave the students a 'pass' or 'not pass' for each rubric category and then gave them specific feedback within each rubric category related to points in the rubric. I think that this really helped me to be calibrated and practiced for reading and grading the writing portfolios.

As Elbow and Belanoff (1986/2009) note, a portfolio system helps anchor grades in the writing classroom to a set of determined standards (p. 99), which is reflected in the faculty assessor comment above. Thus, one of the more encouraging results of changing our process has been the

development of greater synergies between the evaluation of student writing in the classroom and programmatic writing assessment practices.

We view these changes to our FYS program and faculty development workshops as elements in a more equitable writing ecology. By making these changes, we have attempted to address and mitigate the unevenness in the teaching and learning of writing across FYS sections. We see these curricular improvements as having an impact on the writing knowledge students gain and how their writing is assessed through the portfolio. Most simply, the actions faculty are taking to improve their writing pedagogy create opportunities for students to improve their writing (Rutz et al., 2012), and thus students submit portfolios that reflect their improvements. Students also have more agency over how their writing is assessed through revision and reflection. By examining the concerns with our previous assessment practices and aligning our pedagogy to better prepare students for the portfolio, we have instituted a fairer and more equitable structure. We note, however, that the changes we made to our FYS curriculum do not necessarily result in an antiracist assessment process. To move closer to enacting an antiracist assessment through the work we do in our FYS courses, we would need to make antiracist writing assessment explicit in our curriculum; our program would need an antiracist agenda that foregrounds the connections between racism and classroom writing assessment, and indeed to the writing requirements of the college as a whole, such as eliminating a proficiency requirement altogether (Inoue, 2015; Perryman-Clark, 2016; Stewart, 2022). As such, we not only need to pay closer attention to the interconnectedness of the curriculum with the larger assessment process, but we also need to engage FYS faculty in robust discussions of writing instruction and assessment as racialized situations.

Performance Lens

Using a performance lens, we examine whether student identity demographics are variables in our writing assessment results. We borrow the term "performance" from Hammond's (2019) analysis of the "performance talk" in discussions of race and ethnicity in articles published in *Assessing Writing* from 1994-2018, which considers the "differential impacts" of writing assessment practices and outcomes on specific racial demographics (p. 9). Hammond (2019) focuses on how race has been defined and "positioned relative to writing assessment," arguing that "antiracist alternatives for writing assessment" depend on situating our practices within the "broader social meanings" and contexts that shape our agendas (pp. 2–4). In examining racial performance data, Hammond (2019) reminds us that "racial classification[s]" are social constructions, and even choosing who to count and how to count them are embedded in racialized standards (p. 11). In agreement with Hammond, we acknowledge that our definition of "performance" is grounded in our local exigencies and therefore limited. By disaggregating our assessment data—a central recommendation of antiracist scholarship—we examine the racial impacts of our process and consider the underlying biases in our assessment design and institutional context. Recognizing the limitations of our scope, in this section we look at how our writing assessment pass rates are differentiated along three dimensions: race, sex, and first-generation status.[1]

We are guided by the reminder that all assessments "are racial projects, regardless of their purposes . . . intentions, or . . . designs" (Inoue, 2015, p. 52), and so in examining our data to

1 Throughout this article, we use the terms and codes for student demographic information used by our college's Office of Institutional Research. This information is self-reported by students through the Common App then collected and reported by colleges through the Integrated Postsecondary Education Data System, which specifies definitions for race and ethnicity and first-generation students.

understand whether our practices—old and new—resulted in fair and equitable outcomes, we look at how our outcomes intersect with race. We consider both the explicit and implicit inferences of our data because we know that even a well-intended and thoughtfully designed assessment practice is subject to structural biases that result in unequal outcomes for students (Inoue & Poe, 2012; Inoue, 2015). We bring an equity-conscious approach to the analysis of our process and its results, even as we recognize that we have more work to do to address underlying biases and create a fair assessment that moves towards antiracist. Our analysis below focuses mainly on race and ethnicity, but we consider other identity designations with measurable outcomes in our assessment *habitus*.

We are not the first to analyze writing assessment outcomes by demographics (Elliot, 2016; Hammond, 2019; Hamp-Lyons, 1993/2009; Inoue & Poe, 2012; Inoue, 2015; Poe et al., 2018; Stewart, 2022). In their study of undergraduate essays, Roberts, Nardone, and Bridges (2017) ask questions similar to ours about whether assessment rates differed by students' gender and race—although the authors overlook previous scholarship on identity and equity in assessment.[2] We use our disaggregated demographic data to deepen our understanding of the effects of our assessment change and examine whether all students benefited from this change, building on our conclusions in the Curricular Lens section that the portfolio may be a fairer process but not without inequities. The same classroom intervention could result in differential impact on student subpopulations (Eddy & Hogan, 2014), and changes that appear beneficial in the aggregate may mask a minority who did not benefit to the same degree, or worse, were actively harmed by the change. Neither theoretical analysis nor student outcomes alone are sufficient for establishing that our program is equitable, as we recognize that fairness in an assessment ecology is contingent and not "an inherent quality, practice or trait" (Inoue, 2015, p. 56). Rather, they serve as perspectives to examine different parts of the new ecology we created when we changed our assessment.

We ask two questions in this section:

1. Did the portfolio result in increased alignment in pass rates for all students between FYS courses and the program assessment?
2. Do the pass rates differ based on students' race, gender, or first-generation status?

To answer the first, Figure 3 shows the pass rate for both the TWE and portfolio assessments disaggregated by student racial demographics. Due to data limitations, all international students (i.e., any student required to have a visa to study in the U.S.) are listed as U.S. Nonresidents, regardless of country of origin or L2 status. We excluded data from Native Hawaiian/Other Pacific Islanders and students who did not provide their race/ethnicity on the Common App due to small numbers (fewer than two and ten students per year respectively).

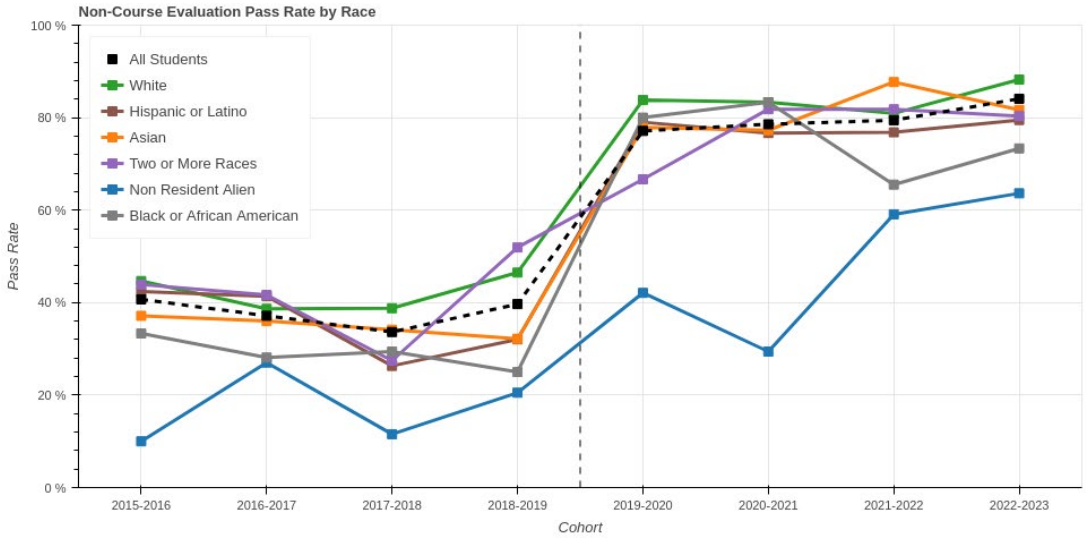
Figure 3 shows that all student groups saw a substantial increase in the pass rate, thus suggesting that the shift to the portfolio benefited all students. We statistically verify this visual understanding with a chi-squared test (see Table 1).

While the pass rates increased, we cannot say that these effects are due to an inherent racial fairness or antiracism of the new system. Moving to our second question, the pass rates for international students (U.S. Nonresidents, the blue line) appear lower than those of other student

2 Roberts et al. (2017) used a multivariate analysis of variance (MANOVA) by taking the average (mean) of four Likert-scale rubric categories. Taking the average (mean) of such scores is generally known to be problematic (Sullivan & Artino, 2013), and thus by extension, the use of MANOVA produces flawed results. These methodological issues reflect the lack of standardization on appropriate statistical tests, an ongoing issue identified in writing assessment (Adler-Kassner & O'Neill, 2011).

Figure 3

Pass Rates of Different Student Racial Groups Before and After Portfolio Assessment



Note. The pass rate is from the TWE only before AY 2019-2020; FYS assessment scores have been omitted.

Table 1

Chi-Square Results for Portfolio Assessment Pass Rates

Racial Designation	*p*	Statistically Significant?
White	< .001	Yes
Hispanic or Latino	< .001	Yes
Asian	< .001	Yes
Two or More Races	< .001	Yes
U.S. Nonresident	< .001	Yes
Black or African American	< .001	Yes

Note. The Chi-square test indicates whether the switch from TWE to portfolio led to differences in pass rates. A statistically significant result means that students had a higher pass rate. This analysis confirms that for all groups, the increase in pass rate after the switch to the portfolio is statistically significant and not due to random chance.

populations, suggesting racialized differences in the pass rates. At the same time, the visually similar pass rates between other student populations do not rule out a lack of racial effects. We examine these questions using a chi-squared test to compare the pass rates of all other student groups to white racialized students, applying the test separately to pass rates for the TWE and the portfolio. The underlying hypothesis is that there would be significant differences in the pass rates for multiple student groups for the TWE, but for fewer student groups for the portfolio, suggesting that the new assessment reduced the effect of race in assessment outcomes and is therefore more equitable. The results for both the TWE and the portfolio are Table 2.

Contrary to our expectations, these results show that pass rates are not different between white racialized students and other student subpopulations, with the exception of U.S. Nonresident students. At our college, about 7% of students are international and, as reported in our data above, coded as U.S. Nonresident students. We can infer that some international students may be L2 speakers and writers who use nonstandard English with writing features that may be flagged as errors by assessors, who may prefer to read standardized prose with stylistic features that meet expectations for "good" writing, whether this preference is explicit or due to unconscious bias. The "writing features" criteria of our rubric are grounded in proficiency expectations that conform to socially dominant conventions of Edited American English (EAE) or Standard Academic English (SAE). Using these features allows FYS faculty and portfolio assessors to evaluate students' writing by the same standards which, on the one hand, promotes a fairer and more equitable methodology. On the other hand, assessing all students by the same language or discourse standards may reify ideological assumptions about privileged dialects and writing traits (Herrington & Stanley, 2012;

Table 2

Chi-Square Results for Pass Rates by Racial Designation

Racial Designation	TWE		Portfolio	
	p	Statistically Significant?	*p*	Statistically Significant?
Hispanic or Latino	0.047	No	0.019	No
Asian	0.023	No	0.288	No
Two or More Races	0.087	No	0.019	No
U.S. Nonresident	< .001	Yes	< .001	Yes
Black or African American	0.012	No	0.031	No

Note. The Chi-square test indicates whether student racial groups had different pass rates than white racialized students. Pass rates between student groups are compared separately for the TWE and for the portfolio. A statistically significant result means that those students had a *lower* pass rate than white students. A non-significant result means that students' race had no effect on the pass rate.

Inoue, 2015; Lippi-Green, 2011). Our assessment is thus subject to a common critique that a rubric, while used to standardize scoring, also enforces prose standards (Balester, 2012). Language standards are inextricable from the racial *habitus* that encourages conformity to white middle- to upper-class social norms (Inoue, 2015, pp. 54–55), and such standardization is a "[racialized] social phenomenon" (Stewart, 2022, p. 3). Our rubric steers assessors towards valuing standardization of grammar, mechanics, and style, which belies the notion of an antiracist (and holistically equitable) assessment practice.

At the same time, the pass rates of students in other student groups are not statistically significantly different from that of white students. Although it is possible that our faculty are more accepting of some kinds of non-SAE features, an alternative hypothesis is that many first-year students "write white," that they have already adopted SAE for academic writing through their prior writing experiences. Under this hypothesis, our results are not due to our faculty assessors' resistance to language hegemony, but a consequence of students adopting "ideologies of assimilation" (Stewart, 2022). Given the difference in pass rates of international students from those of domestic students, this hypothesis is more likely. However, as we have not surveyed faculty attitudes on (non-)SAE writing nor collected data on the prevalence of non-SAE writing features in student portfolios, we lack the nuanced data to conclusively determine the underlying cause of the pattern of pass rates.

While we have improved our assessment mechanism by moving to a portfolio, we recognize that we have retained an acculturationist approach through our rubric (Balester, 2012), which values conformity to SAE and consequently, as shown in our data, results in lower pass rates for international students. Writing features tend to be the "areas where the drive toward standardization is most persistent, visible, and resistant to change" (Balester, 2012, p. 65). We recognize that language standards are used to construct and maintain racial hegemony, and our assessment process exhibits this racialization (Inoue, 2015); we need to further interrogate how our process may use language standards as a means to subordinate students' writing. This disparity in student outcomes reflects a concern that until addressed will prohibit us from moving towards a genuinely equitable assessment process.

Our data points to two other categories in our assessment data and ecology that require more analysis than we provide here: gender and first-generation status. Haswell and Tedesco Haswell's (1996) study of rater bias remains one of the most substantive examinations of the perceived effects of sex/gender on writing performance (the authors use the terms *sex* and *gender* interchangeably). While the study analyzes how assessors may read for gender clues in college-level essays, much of the assessment scholarship on gender focuses on standardized testing of school-aged children. We surmise that one reason there are fewer discrete studies of sex and gender in higher education is because gender effects (to borrow a phrase from Haswell and Tedesco Haswell) are often examined through an intersectional lens that also considers race and ethnicity (Inoue & Poe, 2012; Inoue, 2015). Similarly, because first-generation status intersects with other identity categories including race and ethnicity, there is little scholarship focused on first-generation student outcomes specifically (see Bond, 2019, for a useful review). We summarize our results in Figure 4 and Figure 5, showing that male students and first-generation students consistently have a lower pass rate than female students and non-first-generation students respectively, both before and after the assessment changes. Both differences are statistically significant at $p<0.01$.

Figure 4

Pass Rates for Students by Sex Before and After Portfolio Assessment

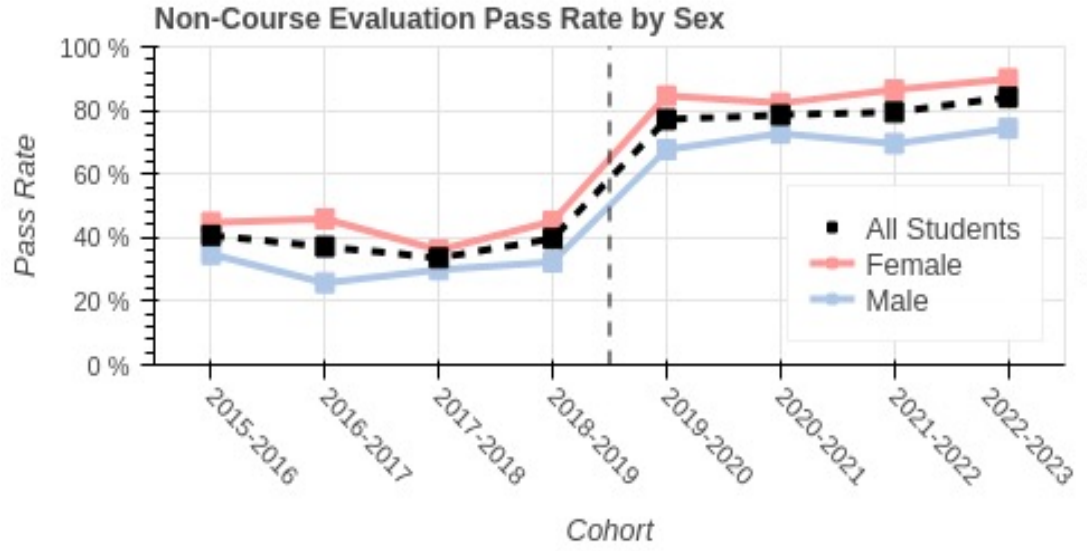
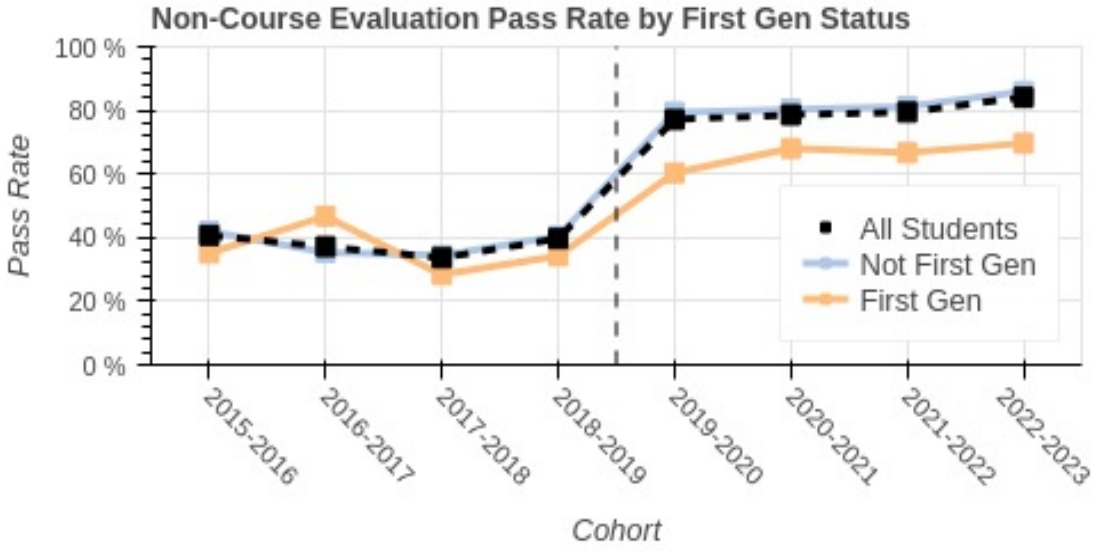


Figure 5

Pass Rates for Students by First-Generation Status Before and After Portfolio Assessment



Overall, our analysis in this section presents mixed results with respect to the equity of our assessment. On the positive side, disaggregating pass rates by race confirms that no one group of students benefited differentially more than another in the shift to the portfolio. We are reassured to see that race did not play a role—at least for domestic students—in the change from the TWE to portfolio assessment, but our analysis does show that international students are passing at a lower rate, and that male and first-generation students are similarly disparately impacted, the causes of which we cannot discern from our existing data. While the portfolio was a step in the direction of a fairer writing assessment process, it is also clear that our assessment is not yet equitable or antiracist. We need to understand the source of the differences in pass rates, which could require further investigation into intersectional factors, students' writing experiences prior to college, and faculty attitudes towards writing. Moving towards an antiracist assessment would also require us to cultivate a community of instructors (and students) interested in examining their classroom practices and experiences in an ongoing assessment of the writing program. Pass rates are a crude measure of student writing, one that is "deeply flawed because it needs a single standard by which to rank students and their performances" (Inoue, 2015, p. 116), but one that faculty are comfortable with; our analysis of disparate impacts might provide the impetus for conversation on how to make our assessment process more equitable.

Reliability Lens

In this section we focus on assessors' evaluations of students' portfolios to analyze reliability as a factor in achieving a fairer, more equitable assessment process. In assessment literature, "reliability" refers to rater consistency across evaluations and "validity," a related term, refers to the argument supported by the adequacy of those assessment decisions (Broad, 1994/2009; Cronbach, 1988; Inoue, 2015; Messick, 1989; Moss, 1994/2009). While portfolio assessment has largely replaced timed exams as a preferred measurement of writing proficiency, researchers have noted challenges in the reliability of portfolio rating and questioned the presumption that reliability is essential to a valid assessment practice (Broad, 1994/2009; Elliot, 2016; Inoue, 2015; Moss, 1994/2009; Nystrand et al., 1993; O'Neill, 2011). Achieving reliability in writing assessment has "focused primarily on getting scorers to agree at an acceptable rate" (O'Neill, 2011, p. 3). Yet, procedures designed to ensure rater consistency or consensus, such as assessor training, may not result in fair judgments of students' writing (Broad 1994/2009; Huot, 1996; Nystrand et al., 1993; Stewart, 2022). Using a reliability lens, we examine rater agreement as one facet of our assessment ecology and question the degree to which achieving equitable outcomes depends on reliability.

By looking critically at our raters' assessments of student portfolios, we examine rater reliability as a way to illuminate the assumptions and consequences of our assessment process (Messick, 1989; Moss, 1994/2009). In a sense, reliability may be at odds with the ideological aims of our assessment; not only does our process not produce entirely equitable results, as we have shown above, we also recognize that our process may assume a connection between reliability and fairness that does not exist (Huot, 1996; Huot & Williamson, 1997/2009; Poe & Inoue, 2016). Evaluative consensus, or interrater reliability, is the "perfect partner" to quantification methods, and trying to reach interrater consensus through an agreed-upon score quantifies writing ability in ways that undercut the potential of portfolio assessment as a fairer process (Broad, 1994/2009, p. 303). Our experience of portfolios versus timed writing exams confirms the benefits of portfolio-

based assessment, as we discuss throughout this article, even as we recognize the problems of scoring portfolios and striving for reliability if we are trying to do something new or innovative.

Our meetings with assessors function in part as "standardizing sessions" (Hamp-Lyons & Condon, 1993/2009, p. 318) designed to explain the goals of the assessment process, ensure that raters understand how to interpret and apply the rubric criteria, and prompt discussion of the varying ways portfolios meet the criteria for passing. Our goal in these sessions is not consensus per se, but achieving some agreement on what constitutes an effective thesis or well-developed analysis of evidence reassures us that assessors are on the same page. We engage assessors in a social or dialogic process that centers their feedback, which scholars have suggested as a way to push back against striving for validity through objectified decision making (Poe & Inoue, 2016). However, our assessors have expressed their need for criteria and standards to guide their portfolio assessment decisions (similar to the readers in Hamp-Lyons and Condon's (1993/2009) study), and so we spend considerable time discussing the rubric and how to use it. As we discuss in the previous sections, we know that assessing student portfolios through a rubric and asking raters to reach their scores using it create the conditions to evaluate students' writing against features of a racialized ideal text (Brannon & Knoblauch, 1982, p. 159; Stewart, 2022, p. 3). Some WPAs see the process of assessment, especially the normative criteria and assessor training used to validate the process, as enacting a form of violence defined as harmful, destructive, and disenfranchising (Lederman & Warwick, 2016). We recognize that the assessor training component of our process is far from enacting a culturally-relevant approach that draws upon the strengths of all student writers and may in itself do unintended harm.

We contextualize our assessment process by looking with nuance at our assessor scoring to identify areas where we might address problems or trends that result in unfair results. In our data below, we consider both the overall percentages of pass-rate agreement between assessors, and we complicate those simpler statistics by analyzing (a sample of) assessor correlations by portfolio criteria. Our assessment analysis allows for a detailed evaluation of where readers agree and disagree, which enables us to focus on patterns in the distribution of the aggregated data and to consider where these patterns suggest assessor biases. We provide a "heatmap" method of data analysis (which we haven't seen before) for examining whether we achieve fairness and equity in our assessment practice.

Table 3 shows the overall pass rate and the agreement percentage between two readers of the same portfolio averaged over all portfolios from 2020 to 2023. These results show variation in

Table 3

Overall Pass Rates by Rubric Category and Assessor Agreement by Category

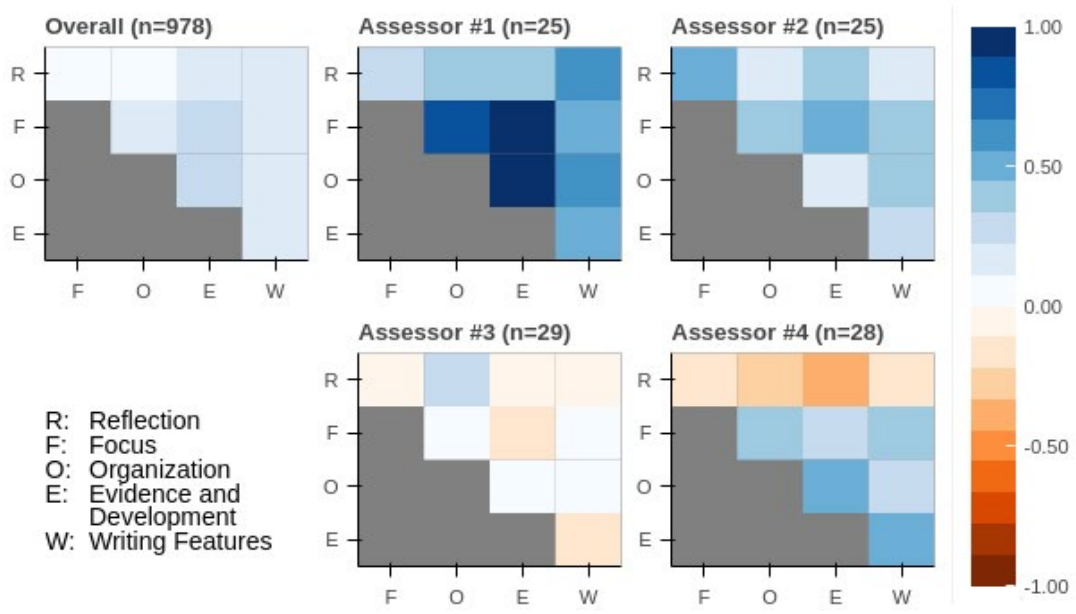
Category	Reflection	Focus	Organization	Evidence & Development	Writing Features
Pass Rate (%)	81.01	69.34	59.25	64.69	71.27
Assessor Agreement (%)	73.18	64.67	58.27	60.74	67.85

assessor agreement between the rubric categories. The category with the highest agreement is reflection at 73.18%, showing that on average, the two readers are aligned; reflection also receives the highest pass rate among the five criteria. Organization is the category with the lowest agreement between assessors, at 58.27% and is also the lowest passing percentage of the criteria. In our process, rater agreement on whether a portfolio meets proficiency does not necessarily correlate with pass rates for each rubric criterion. In other words, two assessors of the same portfolio may reach the same final decision by evaluating the criteria differently. One explanation is because assessors have different understandings of the criteria and biases about the value of each criterion. We also know that some of our assessment criteria may be easier or harder to evaluate than the others. For example, the two criteria with the most agreement (reflection and writing features) are relatively straightforward to assess, as a lack of introspection and grammatical or stylistic errors are easier for assessors to identify. In contrast, the other rubric categories (focus, organization, evidence and development) require deeper and holistic evaluation of the portfolio to determine whether the student has demonstrated those skills proficiently.

Differences between raters extend beyond the individual rubric criteria; there are also meaningful differences in how the criteria relate to each other. While the five criteria are ostensibly independent of each other and have equal weight in the total score (which must be at least 3 of 5 for passing), in practice assessors may see some criteria as more correlated than others. Figure 6 presents heatmaps that show the correlation between each of the rubric categories for the AY 2022-2023 portfolio assessment. The heatmap on the top left, "Overall," shows the correlation of all assessor-portfolio pairs, while the four other heatmaps are the results of four specific assessors. The lower left half of the heatmaps is grayed out, as they duplicate information already in the

Figure 6

Heatmaps of Portfolio Rubric Category Correlations for all Assessors



Note. The heatmaps show portfolio rubric category correlations for all assessors ("Overall" top left) and for four sample individual assessors ("Assessor #1-4").

upper right half. Blue squares indicate positive correlation between two categories (i.e., passing one *increases* the probability of passing the other), while orange squares indicate *negative* correlation (i.e., passing one *decreases* the probability of passing the other). Darker blue or orange indicates stronger positive or negative correlation between categories.

We make several main observations about these results. First, averaging across all assessors, the categories in the rubric are interpreted as mostly independent (weak correlation, as indicated by the lighter shade) but slightly positively correlated (as indicated by the blue color). This pattern matches the intention of the rubric: that each criterion should stand on its own, while acknowledging that influences between the categories are unavoidable. The four heatmaps to the right, however, reveal more substantial differences between individual assessor evaluation of their assigned portfolios—and whether the assessors' results are more or less correlated across the rubric categories. Assessor #1 sees focus, organization, and evidence and development as highly and positively correlated (as shown in the darker blue shading) with writing features only slightly behind. Assessor #2 and Assessor #3 see the categories as mostly independent, with the lighter shading showing some small positive or negative correlations. Finally, Assessor #4 sees focus, organization, evidence and development, and writing features as somewhat correlated, but sees the reflection essay (R) as *negatively correlated* with the other categories; that is, portfolios stronger in focus, organization, and evidence and development may have weaker reflection essays.

These results show the potential impact of assessor evaluation on fair and equitable outcomes for students. Most obviously, in terms of rater reliability, our heatmaps show inconsistency in how assessors interpret the relationship between the rubric criteria, which introduces variance in how students are assessed. In the heatmaps above, Assessor #1 rates focus, organization, and evidence/development as highly correlated, which suggests that a student's "score" on one of these three criteria will likely determine the score on the other two. To put it simply, if a student is randomly assigned this assessor, they will likely pass or fail depending on their proficiency in just one of these three criteria. Conversely, Assessor #3 sees little correlation between the criteria and so a student may fare differently in having this assessor read their writing. While having two assessors read each portfolio aims to reduce reader-subjective scores, our process does not entirely eliminate the possibility of somewhat random and inequitable outcomes. Using a rubric, too, while an attempt to standardize scoring, is not a mechanized process that eliminates the human experience of assessment. Poe and Inoue make a convincing case for "reworking validity theory for the purposes of social justice" by employing a sociocultural approach that understands that assessment decisions—and the arguments drawn from them—are subjective because they are "made from subject positions by people" with their own "particular worldviews, values and dispositions" (Poe & Inoue, 2016, p. 122). Our analysis of the assessor scores suggests that even with "norming" sessions the assessment outcomes and our process overall are influenced by reader bias, which may include anything from assessors' familiarity with higher-order writing concerns, to receptivity to errors, and broader attitudes about students' language abilities.

While we strive to ensure that assessors have a shared understanding of the process, the criteria, and how to apply the criteria to students' writing, we know that there are limitations built into our ecology, as we discussed above. Yet, as we show in the Curricular and Performance sections above, when comparing the TWE to the portfolio, student outcomes *do* show improvement. We have reached a point, then, in our assessment process where we need to reconcile the improvements resulting from our portfolio shift with the broader social and cultural implications

of our assessment work. We might, for example, take up Behm and Miller's (2012) proposition to reject a color-blind assessment process by having discussions with assessors about how "writing assessment practices are constructed within and reinforce a white *habitus*" (p. 136). One way we envision doing so is through the use of our assessment heatmaps as a way to discuss with assessors where our differences in scoring portfolios arise and what these differences might signify. Explicitly discussing with assessors the ways that writing assessment is part of the broader culture of systemic conventions and discourses of power, which may be evident in the patterns or trends we can observe in our heatmaps, is a necessary next step towards creating an antiracist assessment practice. Given the discrepant correlations observed in our heatmaps, we conclude that one of the main elements of reliability—consistency—may be unevenly located in our process. As such, we see reliability as a feature of our assessment process but not solely a way to determine whether our process results in fair and equitable student outcomes.

We also do not consider our faculty assessors through a racial lens, even though we know that the majority white composition of our college's faculty, reflected in our assessment committee, inevitably influences the results of the assessment process and its outcomes (note: our college does not collect faculty demographic information, so we are reporting anecdotal data here). Recognizing the differences in how assessors read portfolios is one step in the dialogue about fair, equitable, and antiracist approaches to the teaching and assessment of student writing. To help validate the fairness of our assessment decisions, we also need to integrate into our training more robust discussion of assessors' perspectives on writing, such as their values and relationship to their work as writing instructors and assessors. As Moss (1994/2009) explains, a critically reflective assessment process should encourage a focus on how assessors see themselves as "constructors of knowledge" with both conscious and tacit "undervalues" about the processes of writing and assessment (p. 157). Such open dialogue with the assessors would be key to a more interrogative assessment process (Hayes & Hatch, 1999; Huot, 1996; Huot & Williamson, 1997/2009; Moss, 1994/2009), and might illuminate assumptions and biases that limit us from achieving our goal of enacting an equitable and antiracist assessment.

Moving Forward: Envisioning an Equitable Writing Assessment

As we have shown above, there is no single metric for measuring the fairness and equity of a writing assessment process. Instead, examining assessment outcomes through various lenses allows us to consider where we meet our goals and identify elements in our process that need more attention. It is important to point out that the umbrella of equity is broader still than our three lenses in this article. We consider briefly the implications of our college-wide writing program and requirements, but could enhance our assessment analysis through an institutional lens examining the intersection of institutional demands (i.e., writing requirements) with our equity initiatives, especially since "decisions about assessment ultimately involve decisions about where to locate power in educational and political institutions" (Huot & Williamson, 1997/2009, p. 334). Since they are directly affected by the assessment, students must be involved in this process more dialogically if we intend to enact a more equitable ecology. Creating a writing assessment is the designing of a system and "design justice requires full inclusion of, accountability to, and ultimately control by people with direct lived experience of the conditions the design team is trying to change" (Costanza-Chock, 2020, p. 99). We have not but could include student perspectives and experiences of the process, especially in our training of assessors, which would enable us to use our

assessment process not only as a means for students to fulfill the college-wide writing requirement but also as a way to "improve teaching and learning" (Moss, 1994/2009, p. 92).

More specifically, while we show with data that our new portfolio assessment is fairer and more equitable, we cannot conclude that the process is equitable from an antiracist standpoint. We have reformed key elements in our process, yet when we bring together and analyze our data through the three lenses we see the limitations in our current process and begin to envision ways we might create an antiracist – and thus more comprehensively equitable – writing assessment. In some ways, our conclusion is not surprising; other scholars reach similar judgments upon reviewing their writing assessment processes. Inoue, for example, in proposing ways to confront the racial effects in assessment, notes the challenges of doing so within institutional structures that validate hegemonic whiteness, socially and in our literacy practices (Inoue, 2015). In their study of the barriers to implementing an antiracist writing pedagogy, Burns, Cream, and Dougherty (2018) similarly focus on the importance of institutional change. They caution that limiting our scope to enacting curricular or programmatic writing assessment processes may leave "intact and undisturbed the institutional ecology of white racial habitus that pervades everywhere else" (Burns et al., 2018, p. 259). What we take from this scholarship and our analysis is not that revisions to writing assessment processes are meaningless, but that we must "embrace action-oriented evaluation" (Burns et al., 2018, p. 286) in our practices to bring about institutional change.

We see the actions we might take to improve our assessment process as generalizable for WAC writing assessment more broadly. First, while our assessment process may not include many features of traditional dynamic criteria mapping, an adapted and collaborative DCM approach might move us—and the field—closer to achieving equitable assessments that are also antiracist. Stewart (2022), for example, proposes an antiracist version of DCM that WPAs might use more intentionally to bring together antiracist writing assessment theory with knowledge of a writing program formed collaboratively with instructors, students, and assessors (Stewart, 2022, p. 7). The process of conversation and negotiation among a program's writing instructors and assessors in determining the evaluation criteria for student writing would not only be valuable for us, but would also be a useful assessment tool for WPAs to consider in representing the writing values within a specific, local, and contextualized institutional community. Another collaborative element of the assessment analysis we offer here, engaged less often in WAC writing assessment scholarship, is a cross-disciplinary approach that relies on and encourages multidimensional disciplinary perspectives. The analysis of the assessment process we offer in this article emerges from our expertise in writing studies/WPA work and computer science; both the methods we employ, such as our heatmap data, and our theoretical frameworks benefit from an interdisciplinary approach to questions of equity and definitions of antiracism, which composition scholars might find useful in assessing for social justice. Finally, even with data showing signs of improvement following our portfolio implementation, we know that any "portfolio-based system of writing assessment must be continually questioned and must continually grow in response to new discoveries and to new phenomena... often engendered by the portfolio evaluation process itself" (Hamp-Lyons & Condon, 1993/2009, p. 316). One point we take from Hamp-Lyons and Condon here is a reminder to analyze our data more frequently to look for new trends or issues that develop over time. Just as we see breaking down the evaluation of assessment into granular metrics to determine where equitable results have been and have yet to be achieved, we see the revisiting of data and findings

as informing broader WAC discussions of the goal of achieving equitable and more socially just assessment.

We hope this paper provides a starting point for enhancing our field's understanding of the facets of equity in a writing assessment process, as well as offering concrete criterion for measuring them. We are not sure that our assessment practice—or any writing assessment practice—will ever be equitable, period, but we hope that by explicitly specifying the different facets of equity we can help the field create writing assessment processes and programs that are fairer, actively reflective, and result in success for all students.

References

Adler-Kassner, L., & O'Neill, P. (2011). *Reframing writing assessment to improve teaching and learning*. Utah State University Press.

Anson, C. (2012). Black holes: Writing across the curriculum, assessment, and gravitational invisibility of race. In A. B. Inoue & M. Poe (Eds.), *Race and writing assessment* (pp. 15–28). Peter Lang.

Balester, V. (2012). How writing rubrics fail: Toward a multicultural model. In A. B. Inoue & M. Poe (Eds.) *Race and writing assessment* (pp. 63–78). Peter Lang.

Behm, N., & Miller, K. D. (2012). Challenging the frameworks of colorblind racism: Why we need a fourth wave of writing assessment scholarship. In A. B. Inoue & M. Poe (Eds.), *Race and writing assessment* (pp. 127–138). Peter Lang.

Bond, C. (2019). 'I need help on many things please': A case study analysis of first-generation college students' use of the writing center. *The Writing Center Journal, 37*(2), 161–194. https://www.jstor.org/stable/26922021

Brannon, L., & Knoblauch, C. H. (1982). Our students' rights to their own texts: A model of teacher response. *College Composition and Communication, 33*(2), 157-166.

Branson, T. S., & Sanchez, J. C. (2021). Programmatic approaches to antiracist writing program policy. *WPA: Writing Program Administration, 44*(3), 71–76.

Broad, B. (2009). 'Portfolio scoring': A contradiction in terms. In B. Huot & P. O'Neill (Eds.), *Assessing writing: A critical sourcebook*. Macmillan. (Reprinted from *New directions in portfolio assessment*, pp. 263–276, by Bob Broad, 1994, Boynton/Cook)

Burch, C. B. (1997). Finding out what's in their heads: Using teaching portfolios to assess English education students and programs. In K. B. Yancey & I. Weiser, *Situating portfolios: Four Perspectives* (pp. 263–277). Utah State University Press.

Burns, M. S., Cream, R., & Dougherty, T. R. (2018). Fired up: Institutional critique, lesson study, and the future of antiracist writing assessment, In M. Poe, A. B. Inoue, & N. Elliot (Eds.), *Writing assessment, social justice, and the advancement of opportunity* (pp. 257–292). The WAC Clearinghouse; University Press of Colorado.

Carter-Tod, S. (2019). Reflecting, expanding, and challenging: A bibliographic exploration of race, gender, ability, language diversity, and sexual orientation and writing program administration. *WPA: Writing Program Administration, 42*(3), 97–105.

Carter-Tod, S., & Sano-Franchini, J. (Eds). (2021). Black lives matter and anti-racist projects in writing program administration [Special issue]. *WPA: Writing Program Administration, 44*(3).

Conference on College Composition and Communication. (2022). Writing assessment: A position statement. *National Council of Teachers of English*. https://cccc.ncte.org/cccc/resources/positions/writingassessment

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need.* The MIT Press.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3–17). Lawrence Erlbaum.

Eddy, S., & Hogan, K (2014). Getting under the hood: How and for whom does increasingcourse structure work? *CBE: Life Sciences Education, 13*, 453–468. https://doi.org/10.1187/cbe.14-03-0050

Elbow, P., & Belanoff, P. (2009). Portfolios as a substitute for proficiency examinations. In B.

Huot & P. O'Neill (Eds.), *Assessing writing: A critical sourcebook*. (Reprinted from "Portfolios as a substitute for proficiency examinations," 1986, *College Composition and Communication, 37*[3], 336–339)

Elliot, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment*, 9(1). https://escholarship.org/uc/item/36t565mm

Elliot, N., Briller, V., & Joshi, K. (2007). Portfolio assessment: Quantification and community. *Journal of Writing Assessment* 3(1), 5–29. https://escholarship.org/uc/item/8nm1m6xc

Gere, A. R., Curzan, A., Hammond, J. W., Hughes, S., Li, R., Moos, A., Smith, K., Van Zanen, K., Wheeler, K. L., & Zanders, C. J. (2021). Communal justicing: Writing assessment, disciplinary infrastructure, and the case for critical language awareness. *College Composition and Communication, 72*(3), 384–412.

Hammond, J. W. (2019). Making our invisible racial agendas visible: Race talk in assessing writing, 1994-2018. *Assessing Writing*, 42. https://doi.org/10.1016/j.asw.2019.100425

Hammond, J. W. (2020). Toward a social justice historiography for writing assessment. In M. Poe, A. B. Inoue, & N. Elliot (Eds.), *Writing assessment, social justice, and the advancement of opportunity* (pp. 41–70). The WAC Clearinghouse; University Press of Colorado.

Hamp-Lyons, L., & Condon, W. (2009). Questioning assumptions about portfolio-based assessment. In B. Huot & P. O'Neill (Eds.), *Assessing writing: A critical sourcebook* (pp. 315–329). (Reprinted from "Questioning assumptions about portfolio-based assessment," 1993, *College Composition and Communication, 44*[2], 176–190)

Haswell, R. H., & Tedesco Haswell, J. (1996). Gender bias and critique of student writing. *Assessing Writing, 3*(1), 31–83. https://doi.org/10.1016/S1075-2935(96)90004-5

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication, 16*(3), 354–367.

Herrington, A., & Stanley, S. (2012). CriterionSM: Promoting the standard. In A. B. Inoue & M. Poe (Eds.), *Race and writing assessment* (pp. 47–62) Peter Lang.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*(4), 549–566. https://www.jstor.org/stable/358601

Huot, B., & Williamson, M. M. (2009). Rethinking portfolios for evaluating writing: Issues of assessment and power. In B. Huot & P. O'Neill (Eds.), *Assessing writing: A critical sourcebook*, 330–342. (Reprinted from *Situating portfolios: Four perspectives*, pp. 43-56, by K. B. Yancey & I. Weiser (Eds.), 1997, Utah State University Press)

Inoue, A. B., & Poe, M., (Eds.) (2012). *Race and writing assessment*. Peter Lang.

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: teaching and assessing writing for a socially just future*. The WAC Clearinghouse; Parlor Press.

Lederman, J., & Warwick, N. (2016). The violence of assessment: Social (in)justice and the role of validation. In M. Poe, A. B. Inoue, & N. Elliot (Eds.), *Writing assessment, social justice, and the advancement of opportunity* (pp. 229-255) The WAC Clearinghouse; University Press of Colorado.

Lau, A. (2013). Timed writing assessment as a measure of writing ability: A qualitative study. *Discussions,* 9(2).

Lippi-Green, R. (2011). *English with an accent: Language, ideology, and discrimination in the United States*. Routledge. https://doi.org/10.4324/9780203348802

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13– 103).

Moss, P.A. (2009). Can there be validity without reliability? In B. Huot & P. O'Neill (Eds.), *Assessing writing: A critical sourcebook*. (Reprinted from "Can there be validity without reliability?," 1994, *Educational Researcher, 23*[2], 5–12)

Nystrand, M., Cohen, A. S., & Dowling, N. M. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment, 1*(1), 53–70. https://doi.org/10.1207/s15326977ea0101_4

O'Neill, P. (2011). Reframing reliability for writing assessment. *Journal of Writing Assessment*, *4*(1). https://escholarship.org/uc/item/6w87j2wp

Petersen, J. (2009). 'This test makes no freaking sense': Criticism, confusion, and frustration in timed writing. *Assessing Writing, 14*(3), 178–193. https://doi.org/10.1016/j.asw.2009.09.006

Perryman-Clark, S. M. (2016). Who we are(n't) assessing: Racializing language and writing assessment in writing program administration. *College English, 79*(2), 206–211. https://www.jstor.org/stable/44805919

Poe, M., & Inoue, A.B. (2016). Toward writing as social justice: an idea whose time has come. *College English, 79*(2), 119-126.

Poe, M., Inoue, A. B., & Elliot, N. (Eds.). (2018). *Writing assessment, social justice, and the advancement of opportunity*. The WAC Clearinghouse; University Press of Colorado. https://doi.org/10.37514/PER-B.2018.0155

Prendergast, C. (1998). Race: The absent presence in composition studies. *College Composition and Communication, 50*(1), 36–53.

Roberts, J., Nardone, C. F., & Bridges, B. (2017). Examining differences in student writing proficiency as a function of student race and gender. *Research and Practice in Assessment*, *12*, 59–68.

Rutz, C., Condon, W., Iverson, E. R., Manduca, C. A., & Willett, G. (2012). Faculty professional development and student learning: What is the relationship? *Change: The magazine of higher learning, 44*(3), 40–47. DOI: 10.1080/00091383.2012.672915

Stewart, M. K. (2022). Confronting the ideologies of assimilation and neutrality in writing program assessment through antiracist dynamic criteria mapping. *Journal of Writing Assessment*, *15*(1). https://escholarship.org/uc/item/7rq4n47t

Sullivan, G. M., &, Artino, Jr., A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education, 5*(4), 541–542. DOI: 10.4300/JGME-5-4-18

Trimbur, J. (1996). Why do we test writing? In E. M. White, W .D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, and practices* (pp. 45-48). Modern Language Association of America.

Weiser, I. (1997). Revising our practices: How portfolios help teachers learn. *Situating portfolios: Four perspectives*, 293-305.

White, E. (1994). *Teaching and assessing writing*. Jossey Bass.

Yancey, K. B., & Weiser, I. (1997). *Situating portfolios: An introduction*. In K. B. Yancey & I. Weiser (Eds.), *Situating portfolios: Four perspectives* (pp. 1–18). Utah State University Press.