Journal of Writing Assessment

Queen, B., Kirby, K., Eslami, M., & Denaro, K. (2025). Exploring fairness and seeking social justice for writing assessment: ePortfolios, language difference, and metacognition. *Journal of Writing Assessment*, *18*(2). **DOI**: 10.5070/W4.jwa.1599

Exploring Fairness and Seeking Social Justice for Writing Assessment

ePortfolios, Language Difference, and Metacognition

Bradley Queen, University of California, Irvine, US, bradley.queen@uci.edu
Kate Kirby, University of California, Irvine, US, kathark@uci.edu
Maryam Eslami, University of California, Irvine, US, eslamim@uci.edu
Kameryn Denaro, University of California, Irvine, US, kdenaro@uci.edu

Abstract: This quantitative validation analysis applies antiracist methods to longitudinal ePortfolio assessment data to study language difference through the lens of a metacognitive literacy construct. With interdisciplinary research reshaping the field of writing assessment using quantitative and intersectional demographic approaches, this essay advances language difference as a meaningful register of validity evidence and an indicator of fairness across linguistically and racially heterogeneous students sorted into three cohorts that establish comparisons of ePortfolio assessment scores from samples tracking from 2016–2020. To contribute to the critical study of social justice in writing assessment, this exploratory analysis offers nuanced responses to its guiding heuristic question: Can ePortfolios be instruments of fairness in a local assessment ecology? For this formative curricular assessment, rigorous statistical methods complicate claims derived from the ePortfolio assessment results, with post-hoc power calculations and disparate impact analysis used to search for differences between language cohorts and intersectional demographics defined by race/ethnicity, socioeconomic status, and first generation. These quantitative methods further inferences about the ePortfolio instrument's fairness by problematizing the use of singular demographic aggregations for underrepresented students when attempting to validate assessment constructs and engage in the ongoing study of fairness and social justice in a local writing assessment ecology.

Keywords: ePortfolios, writing assessment, formative curricular assessment, fairness, language difference, antiracist writing assessment, validity, reliability, power analysis, disparate impact, intersectionality

Introduction

As an empirical study of fairness that applies quantitative methods to the assessment of ePortfolios to study language difference through the lens of a metacognitive literacy construct, this validation analysis derives its warrants from adjacent fields currently recasting the interdisciplinary area of writing assessment as a progressive project for the critical study of social justice. With the reconceiving of fairness in the *Standards for Educational and Psychological Testing*, which casts it as "a fundamental validity issue" (American Educational Research Association [AERA] et al., 2014, p. 49), fairness is now recognized alongside reliability/precision and validity as a third pillar of ethical standards used to scrutinize assessment and instrument designs, measurement procedures, and attendant consequences for stakeholders (Elliot, 2015, 2016; Poe & Inoue, 2016). Building on developments in validity theory by Messick (1989) and Kane (2006), in which the unified theory of validity forwards the interpretation-use conception of validity as evidence-driven argument, and drawing from Mislevy's (2018) situating of validation research within a multidisciplinary sociocognitive frame, researchers in writing studies and adjacent educational measurement fields are currently engaged in the search for theories and methods that might enact social justice writing assessments and clarify the meaning of fairness within assessment ecologies (Cushman, 2016; Driscoll & Zhang, 2022; Elliot, 2016; Inoue, 2015; Kelly-Riley, 2011; Kelly-Riley & Whithaus, 2016; Kelly-Riley et al., 2016; Lederman, 2023; Poe & Cogan, 2015; Poe & Elliot, 2019; Randall et al., 2024a; Randall et al., 2024b).

One strand of research of importance to this study uses culturally responsive models of quantitative writing assessment to study the social construction of race/ethnicity and intersectionality through literacy and language assessment and to search for evidence of bias in assessment ecologies (Inoue, 2015; Poe et al., 2023; Randall et al., 2024b). This critical project has been influenced by a multifaceted body of research. Literacy assessments have long been seen as sites of disparate racial/ethnic impacts in terms of educational attainments and long-term socioeconomic trajectories (Bowles & Gintis, 2002; Hamp-Lyons & Condon, 2000; Haswell & Elliot, 2019; Inoue & Poe, 2012; Poe et al., 2014; Raudenbush & Kasim, 1998). Moreover, a long-span interdisciplinary engagement with the problematic of language difference tracks from the influential CCCC (1974) white paper, "Students' Right to Their Own Language (SRTOL)," which set the field to grapple with language difference and injustice (Gere et al., 2021; Perryman-Clark et al., 2014) through to the advocacy of second language scholars urging mainstream writing studies to "embrace language difference as the new norm" (Matsuda, 2006, p. 648) and assimilate translingualism's critical politics of student agency and the ecological dynamics of literate resources (Canagarajah, 2013; Horner & Trimbur, 2002; Horner et al., 2011a; Horner et al., 2011b; Lee, 2018; Lu & Horner, 2013).

These perspectives motivate what this study identifies as an enregisterment approach to validation analysis in writing assessment in which rigorous quantitative engagement with language difference might indicate bias along underrepresented intersectional demographics and shed light on fairness within an assessment ecology. Enregisterment, a concept borrowed from linguistic anthropology, theorizes that linguistic formations become indexes of social meaning within specific ecologies/social systems by signifying other meaningful registers of personhood and identity such as race/ethnicity, and first generation and socioeconomic status (Rhodes et al., 2020; Rosa, 2019; Urciuoli, 2009).

With justice as the "ultimate aim of assessment" (Poe et al., 2015, p. 5), and driven by the principle that "the context of injustice cannot be disaggregated from (or ignored by or through) the assessment experience" (Randall et al., 2024b, p. 3), recent studies forward racial validity (Inoue, 2009; Inoue & Poe, 2012) and justice-oriented, antiracist validation (JAV) (Randall et al., 2022; Randall et al., 2024a) as approaches to the study of fairness in writing assessment that acknowledge histories of discrimination in literacy assessments (Elliot, 2016). Inoue (2015), for example, ties racial validity to fairness as a heuristic that searches for bias in assessment ecologies by locating entrenched logics of discrimination and attendant assessment practices that disparately impact multilingual students and underrepresented minorities. Guided by the assumption that language education and assessment are rooted in histories of coloniality and "the structural white *habitus* that make institutions hostile to diverse educational communities" (Poe et al., 2018, p. 19; see also Behm & Miller, 2012; Che, 2022; Inoue, 2015), Randall et al. (2024b) use quantitative critical race theory (QuantCrit) as they build on Messick's, Kane's, and Mislevy's ideas to forward JAV as an approach to validation analysis that assumes the empirical and ideological inertia of oppressive systems influences formations of institutionally collected data and that the move to use such data to problematize such systems casts aside easy claims of neutrality and objectivity for quantitative analysis. QuantCrit weaves together critical race theory's concerns for the absence of justice for historically oppressed people within legal and social systems otherwise keyed to color-blind abstractions of justice with intersectional quantitative and demographic approaches to the analysis of institutional data (Randall et al., 2024a). QuantCrit therefore suggests that rigorous empirical methods for writing and literacy assessment should embed themselves within the inertial flow of institutional data to engage in recursive study of institutional systems and to use validation analysis as an ongoing heuristic to assess fairness within postsecondary literacy assessment ecologies.

Taking cues from these developments, this exploratory validity study uses quantitative methods to address a humanist question that serves as a heuristic for ongoing curricular attunement: Can ePortfolios be instruments of fairness in a local assessment ecology? Evidence is gathered from three ePortfolio assessments with samples tracking from 2016–2020. The study at hand draws from a dense body of longitudinal assessment data to derive validity inferences about the ePortfolio instrument's fairness across linguistically heterogeneous students in a capstone general education research-writing course at a public R1 university. This empirical foundation is established using a linguistic demographic variable that defines the sample sets for each of the three writing assessments, which took place in the summers of 2017 (n=630), 2018 (n=370), and 2020 (n=354). This variable, a self-reported primary home language register, defines three cohorts of students—Monolingual English (MoE), Multilingual English (MuE), and Multilingual (Mu)—as the basis for comparisons of assessment results. With each portfolio receiving scores from two different readers, inter-reader reliability estimates suggest how reliably readers carried out their acts of judgement from within common frames of value-laden information with coefficients that register the statistical degree of the strength of shared interpretations of the assessment method, the rubric and its seven elements, and of the literacy construct under study—defined as rhetorical metacognition, an ecological construct contoured by curricular opportunities for self-reflection and self-monitoring and evaluation. In layering fashion, subsequent analyses enact quantitative antiracist methods by complicating inferences derived from the assessment results using post hoc power calculations and disparate impact analysis. Post-hoc power analysis furthers probabilistic evidence used to determine whether the assessment methods detected hypothetical differences

among assessment cohorts, and disparate impact analysis searches for patterns of difference for underrepresented students using mixed effects linear regression modeling to reach intersectional demographics and to offer nuanced responses to the study's guiding question.

Literature Review

The reconceiving of fairness in *The Standards* (AERA et al., 2014) accelerated the social justice turn in writing assessment. Kane's (2006) and Messick's (1989) innovative contributions to the integrated theory of validity positioned construct validity as subsuming other forms of validity. Messick (1989) defined the unified model 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" (p. 15). Inoue's (2009) conception of racial validity recast the unified model and its evidence-driven framework as an "argument that explains the degree to which empirical evidence of racial formations around an assessment and the theoretical frameworks that account for racial formations support the adequacy and appropriateness of inferences and decisions made from the assessment" (p. 110). Subsequently, with the emergence of fairness as an integrative concept in *The Standards* (AERA et al., 2014), writing studies and assessment researchers leaned into longstanding sociocognitive assumptions in empirical research to develop critical validation models that reach toward ecological approaches to writing assessment by way of quantitative and demographic methods (Inoue & Poe, 2012; Inoue, 2015; Kelly-Riley, 2011; Poe & Elliot, 2019; Poe & Inoue, 2016; Poe at. al., 2015; Randall et al., 2024a; Randall et al., 2024b).

The theoretical development of fairness deepens in writing studies with the publication of Elliot's (2016) essay, "A Theory of Ethics for Writing Assessment." Drawing from John Rawls and Robert Merton, Elliot recasts Rawls' difference principle about redistributive justice and pulls from Merton's thinking on social structural organization and social deviance to establish assumptions for a theory of fairness: all are assumed to have the same indefeasible claims to basic liberties, and inequalities exist only to provide equal opportunities for access to public goods and to benefit the least advantaged members of a social system. Fairness in writing assessment, then, is defined by Elliot (2016) "as the identification of opportunity structures created through maximum construct representation. Constraint of the writing construct is to be tolerated only to the extent to which benefits are realized for the least advantaged" (p. 6). This conception of construct representation also draws from Messick (1989) and Kane (2006) and forwards a description of the construct of writing to mitigate inequities that have historically resulted from narrow construct representations, as evidenced by the histories of standardized testing and timed writing placement exams for postsecondary students. Elliot's definition comprises both agentive and ecological aspects, with the former described by four domains—cognitive, interpersonal, intrapersonal, and physiologic—and the latter by conceiving of writing as sociocognitive and complexly contextualized. Elliot's (2016) formative conceiving of fairness for writing assessment positions the critical study of fairness as the objective of the technical process of validation.

Notwithstanding the recasting of fairness in the *Standards* (AERA et al., 2014) as a foundational measurement concept alongside validity and reliability/precision, current thinking argues that the *Standards* has not "sufficiently addressed justice as primary virtue of social institutions" (Poe et al., 2023, p. 194; see also Randall et al., 2024a). To move validation theory forward, current research interprets "fairness as justice," keying on Rawls (Randall et al., 2024a, p.

208), and therefore deduces that if justice is the objective of validation analysis, any study should demonstrate through statistical and quantitative methods and nuanced critical analysis whether an assessment instrument attends to a local conception of fairness. Poe et al. (2023) propose "racial justice extensions" to the *Standards* to engage in principled theorizing of validation approaches that redress historical inequities in educational measurement and account with rigorous deliberation for cultural and linguistic difference in assessment design and interpretation (p. 194). They offer four theoretical frames through which to view potential approaches to social justice validation: intersectionality theory, moral philosophy, disparate impact analysis from legal fields, and social theories of learning and social situatedeness from the field of education.

While noting "a dearth of intersectional quantitative studies in the field of education, and fewer still in the assessment/measurement literature," Randall et al. (2024b) deploy QuantCrit as an approach to understanding bias and injustice in assessment ecologies (p. 12). Recognizing that quantitative validation analysis should be an important heuristic for assessment and that the posing of questions about fairness and social justice is intended to yield nuanced inferences for on-going study, Randall et al. (2024b) focus on conventional assessment areas to forward a critical JAV framework: construct articulation and validation, data analysis, and data interpretation/score reporting. Drawing from Mislevy who argues that validation analysis should "explore the interplay between model-based reasoning and the linguistic, cultural, and substantive patterns that shape tasks, individual performances, and interpretations of both" (Randall et al., 2024b, p. 8), the authors theorize QuantCrit as motivating JAV methods that interrogate the value-laden cultural and social systems that tacitly underlie assessment designs and attendant validation studies by applying intersectionality as a key analytic consideration.

If the objective of validation analysis is to assess fairness, then attempting to account for systemic bias necessitates the use of methods that can reach across a wide range of intersecting identity formations and seek out disparate impacts tied empirically to assessment instruments, their outcomes, and the local assessment ecologies that define the values, proficiencies, and literacies being assessed. QuantCrit brings intersectionality to the fore in a burgeoning research effort seeking to understand fairness ecologically (Poe et al., 2018; Wardle & Roozen, 2012) by forwarding analytical methods such as disparate impact analysis (Poe et. al, 2014) as a validation tool "for understanding the local effects of writing assessment on diverse groups of students" (Kelly-Riley & Elliot, 2016, 2021; Kelly-Riley et al., 2016; Poe et al., 2014, p. 288; Poe et al., 2014; Poe & Cogan, 2016).

Catalyzed by these developments, assessment researchers are engaged in the search for justice and for clarity about fairness itself while theorizing validation analysis as an ongoing heuristic to assess fairness within postsecondary literacy assessment ecologies, even if this critical project is not yet thoroughly described by empirical and theoretical research (Kelly-Riley & Elliot, 2021; Kelly-Riley et al., 2016). The study presented in this essay uses disparate impact methods to search for statistically significant longitudinal differences among three linguistic cohorts across three ePortfolio assessments and the metacognitive literacy construct they measure. ePortfolio research in writing assessment, as noted by several studies, is a burgeoning area with substantial gaps in quantitative demographic analysis of these construct-rich assessment instruments (Bryant & Chittum, 2013; Kelly-Riley et al., 2016). The 2016 study by Kelly-Riley, Elliot, and Rudniy notes the slow development of portfolio research, as this study is the first to gather evidence from ePortfolio scoring and then analyze assessment results by validating demographic data to draw

inferences about the fairness of assessment outcomes. Taking methodological cues from the 2014 revision to *The Standards*, Kelly-Riley et al. (2016) gather reliability evidence to inform validity inferences to reach claims about the fairness of the assessment ecology in their study.

Portfolios have long been seen as the most meaningful instrument in writing studies for the study of authentic literacy performances. As ecologically enmeshed (Yancey, 2009), portfolios have enabled researchers to problematize psychometrics and move toward hermeneutic approaches for writing assessment (Broad, 2003; Gipps, 1999; Moss, 1994; Smagorinsky, 2006) and have catalyzed the long turn toward validity as a guiding concept for writing assessment. With their interdisciplinary sociocognitive roots, portfolios offer richly contextualized interpretations of literate abilities and learning objectives as they nurture rhetorical consciousness (Hamp-Lyons & Condon, 2000), document the values and the ecological contours of their creation and assessment (Black et al., 1994), and inculcate student-centered approaches that motivate self-reflection and metacognition. (White, 2005; Yancey, 1996).

Moving beyond the Elbow-Belanoff portfolio model and its focus on literate abilities gleaned from the assessment of multiple examples of student writing, Yancey (1998, 1999) posits four types of knowledge students can acquire through portfolio pedagogy: reflective/metacognitive self-awareness about habits and practices, content knowledge, task knowledge that includes rhetorical awareness, and knowledge of themselves as agents. Multidisciplinary research has long argued that metacognition is essential for students to regulate adaptive responses across different learning contexts (Bawarshi, 2017; Driscoll & Zhang, 2022; Gorzelsky et al., 2017; Negretti, 2012; Scott & Levy, 2013; Taczak & Robertson, 2017), particularly when considering that metacognition is the ability to reflect on and use knowledge that is both new and accumulated (Negretti, 2012; Serra & Metcalfe, 2009). Theories of metacognition's efficacy as a fundamental disposition for learning derive from studies in postsecondary writing contexts, second language and English for academic purposes fields, and special education contexts (Beaufort, 2007; Negretti & Kuteeva, 2011; Nowacek, 2011; Salomon & Perkins, 1989; Yancey et al., 2014).

Negretti (2012), for example, uses a participatory constructivist method to document how academic writers in second language contexts use metacognition to regulate their performances, a method that complicates traditional approaches to language acquisition in which students study the grammatical features available in academic writing situations. Moreover, signaling a broader shift in L2 contexts, one similar to the evolution of thought within writing studies—from "positivist to constructivist perspectives of language assessment in general and writing assessment in particular" (Lam, 2017, p. 87)—portfolio assessment methods with metacognitive focal points have taken root "within a global assessment reform movement, which emphasizes using assessment to promote self-regulated learning and improve pedagogy with a focus on empowering" students (Lam, 2017, p. 87). Nevertheless, Lam (2017) notes the inadequacies of portfolio assessment research in both second- and first-language writing classrooms, criticizing portfolio systems for training reflection on task compliance, which minimizes student agency.

Within the "Framework for Success in Postsecondary Writing" (CWPA et al., 2014), a guiding document for writing studies, metacognition is situated as both a habit of mind and a diffuse disposition underlying other habits and experiences, an articulation suggesting that metacognition is a core disciplinary concept and that further theorizing is needed to pursue clarification of it as an object of research (Portanova et al., 2017). One recent study conceives of metacognition as a mediating factor between personal and social ecologies in a longitudinal study of two writers that

tracks them from colleges to workplaces and advanced study experiences. Driscoll and Zhang (2022) assemble qualitative data to describe critical interactions among these ecologies to further substantiate established conceptions of metacognition (sec. 1; see also Taczak & Robertson, 2017): that metacognition, defined as both awareness and regulation, can vary from person to person and can lead to long-term growth for students taught to recognize idiosyncratic characteristics that can lead to negative knowledge transfer and/or those that can foster growth and advanced learning. Another recent study (Gorzelsky et al., 2017) offers the concept of constructive metacognition, noting that despite its presence in reputable scholarship (Wardle, 2009), metacognition has long been fuzzy in writing studies. The authors argue that "none . . . have identified the *metacognitive (sub)components* and how they operate in writing" (Gorzelsky et al., 2017, p. 219). Drawing from the work of Scott and Levy (2013), Georghiades (2004), and Yancey (1998), Gorzelsky et al. (2017) define constructive metacognition as "a critically reflective stance likely to support transfer of writing knowledge across contexts" (p. 216). This research breaks new ground methodologically in reaching across institutional contexts to define constructive metacognition and by building on previous constructivist research (Beaufort, 2007; Yancey et al., 2014), which has been instrumental to understanding reflection and student agency using small case studies.

Methods

Institutional Data and Language Difference

Building on these developments, this exploratory study of fairness uses a linguistic demographic variable to create three cohorts compared across three ePortfolio assessments: 2017 (n=630), 2018 (n=370), and 2020 (n=354). Collected for institutional statistics across the University of California system in response to demographic questions on admissions applications, the "primary home language" variable enregisters other intersectional demographics such as race/ethnicity, first generation, and socioeconomic status as it defines the three assessment cohorts to study language difference, search for biases, and assess fairness: monolingual English students (MoE) for whom English is the only language spoken at home, multilingual English students (MuE) for whom English and another language or languages are primary, and multilingual students (Mu) with primary languages other than English spoken at home.

As an institutional statistic for a linguistically heterogenous student body in which upwards of 75% of the total student population is multilingual, the "primary home language variable" is not manifestly "colorblind" (Davila, 2017; see also Behm & Miller, 2012; Stewart, 2022) and therefore does not lend itself to the easy dispensing of difference in assessment design. Problematizing the use of conventional typologies to generate comparisons among racial/ethnic groups, it sorts instead by primary home language in which intersectional demographics are subsumed and indexed by language identification (see Tables 1 and 2). Table 2 documents the language variable enregistering races/ethnicities and suggests that it may therefore elide critical indicators of difference when assessing performative literacies if the analysis of intersectionality demographics is not reached by quantitative methods. However, this foundational linguistic variable and the rigorous QuantCrit methods used here to study fairness critically enable the assessment of its validity by demonstrating that it is a telling empirical indicator of intersectional differences within a public postsecondary writing assessment ecology, the site at which literacy biases and discriminatory practices have historically been replicated.

Table 1

ePortfolio Assessments, Participant Characteristics: First (2017), Second (2018), and Third (2020) Assessments

Variable	Assessment Year							
	All Years Combined (*N*=1351)		Year 1 (*N*=630)		Year 2 (*N*=367)		Year 3 (*N*=354)	
	n	%	*n*	%	*n*	%	*n*	%
First language								
MoE*	368	27.2	149	23.7	132	36.0	87	24.6
MuE	492	36.4	206	32.7	147	40.1	139	39.3
Mu	491	36.3	275	43.7	88	24.0	128	36.2
Low income								
No	900	66.6	412	65.4	244	66.5	244	68.9
Yes	451	33.4	218	34.6	123	33.5	110	31.1
First Gen								
No	674	49.8	286	45.4	199	55.4	189	53.4
Yes	669	49.5	344	54.6	160	44.6	165	46.6
Race/Ethnicity								
African American	43	3.2	23	3.7	14	3.8	6	1.7
Asian	665	49.2	309	49.0	210	57.2	146	41.2
White	159	11.8	76	12.1	43	11.7	40	11.3
Hispanic	335	24.8	169	26.8	81	22.1	85	24.0
Other Race+	135	10.0	39	6.2	19	5.2	77	21.8
Decline to State	14	1.0	14	2.2	0	0.0	0	0.0
Missing	8	0.7						

* Monolingual English students (MoE), multilingual English students (MuE), and multilingual students (Mu).

+ The *other race* category includes American Indian, Alaskan Native, Native Hawaiian or Other Pacific Islander.

Table 2

Crosstabulations, Race/Ethnicity & Primary Home Language, ePortfolio Assessments (2017, 2018, 2020)

Race/Ethnicity	Language*			Totals	
	MoE	MuE	Mu	*n*	%
Hispanic	69	151	115	335	24.8
Asian	129	269	267	665	49.2
White	99	36	24	159	11.8
African American	31	10	2	43	3.2
Other Race+	34	24	7	135	10.0
Decline to State	6	2	6	14	1.0
Totals					
n	368	492	491	1,351	
%	27.2	36.4	36.4		100

* Monolingual English students (MoE), multilingual English students (MuE), and multilingual students (Mu).

\+ The *other race* category includes American Indian, Alaskan Native, Native Hawaiian or Other Pacific Islander.

Validity Evidence

Evidence to support validity inferences is situated into four types defined by Kane (2006) and Poe et al. (2014): scoring, generalization, extrapolation, and consequential. Scoring evidence derives from the longitudinal analysis and statistical significance testing of ePortfolio assessment results and from two forms of inter-reader reliability that support interpretations of scoring (see Table 3 in Supplemental Materials). Two types of sources serve as generalization evidence. For one, the two inter-reader reliability analyses register the statistical degree of the strength of shared interpretations of the assessment method, the rubric and its seven elements, and of the metacognitive writing construct. For the second, post-hoc power calculations apply to each assessment sample and support claims that each assessment cohort represents the broader populations from which they were randomly selected. Extrapolation evidence extends interpretations of the assessment data beyond the immediate assessment context and is sourced from the inferential synthesis of scoring evidence, generalization evidence, and disparate impact analysis that uses mixed effects linear regression to search for unintended consequences along intersectional demographics (see Tables 4, 5, 6, and 7 in Supplemental Materials). Together, these sources of validity evidence offer support to claims about the fairness of the writing construct and the ePortfolio instrument and provide sources of consequential evidence in the form of responses to this study's exploratory question.

Assessment Procedures

Assessment procedures are presented in the order they were applied. Three ePortfolio assessments were undertaken in the summers of 2017, 2018, and 2020. For each of the three cohorts—monolingual English students (MoE), multilingual English students (MuE), and multilingual students (Mu)—mean scores for each rubric area establish the basis for comparisons among them and are used to derive statistical significance indicators. Linear mixed effects models were used. The dependent variable is represented by the possible values for each rubric question treated on the continuous scale, and the independent variable is the linguistic cohort. Likelihood ratio tests that compare models with and without adjustment for language cohort were used to determine whether each language cohort explained a statistically significant amount of variation documented by the mean scores (see Table 3).

Portfolio samples for each assessment come from regular academic year quarters. ePortfolios were distributed to raters after being randomly selected from among ePortfolios collected across the academic quarters preceding the summer assessments. The third assessment includes samples from the first full quarter of remote instruction due to the COVID-19 quarantine. Demographic stratifying was performed on the sample for the first assessment but not for the two subsequent assessments.

The assessment rubric uses a five-point Likert scale—excellent (5), high average (4), average (3), low average (2), very poor (1)—and forwards an assessment method similar to the summative grading method used by teachers in their classrooms. It first asks readers to develop an initial holistic impression by reading through an ePortfolio's reflective introduction—a multimodal composition of up to 1,300 words, which is the guiding document for each portfolio—before perusing the rest of the portfolio and its learning artifacts to assess the persuasiveness of the metacognitive arguments made by students about their learning. Next, readers assess five traits, and then they give a second holistic rating that may adjust the first holistic impression.

All readers for each assessment had experience teaching the research writing curriculum from which the assessment samples were drawn. The first assessment had ten readers, the second had eight, and the third had ten readers. Four of the same raters participated in assessments one (2017) and two (2018), while two readers did the second and third (2020), and one reader participated in all three. In sum, twenty-one different teachers participated, with Graduate Teaching Assistants and Lecturers splitting the overall pool of twenty-eight readers in half with fourteen apiece. All readers participated in a three-hour calibration session prior to rating their samples remotely.

Inter-rater Reliability Estimates

Two types of inter-rater reliability estimates—weighted quadratic kappa and Spearman's rho—were calculated for each rubric question across the three assessments to determine whether and to what degree ePortfolios were read reliably (see Table 3). The kappa statistic is a traditional consensus estimate, indicating the strength of exact agreement among raters in their scoring by mathematically weighting all differences equally outside of exact agreements. Spearman's rho gives a consistency estimate, indicating the strength of agreement as consistent patterns of judgement (Stemler, 2019) whether they are close or far apart. The weighted quadratic form of the Kappa statistic is used to allow for more nuanced estimates of consensus agreement. White et al. (2015) suggest reliability scales for writing portfolios (see Table 8) in which the tiers of strength for

Table 8

The Whale Scale (White et al., 2015)

	Weighted Kappa (non-adjudicated)	Spearman rho (non-adjudicated)
High	.46-.69	.48-.71
Medium	.23-.45	.23-.47
Low	.1-.22	.1-.22

Table 9

Post-Hoc Power Analysis

Assessment Year	Sample Size*	Multivariate Tests				Power Analysis				
		Pillai's Trace	F	df	*p*	Alpha Level	Critical F	Pillai's Trace (Pillai V)	Effect Size $F^2(V)$	Power
Year 1 (2017)	630	0.24	11.9	14, 1244	<.001***	0.05	1.70	0.24	0.13	0.99
Year 2 (2018)	367	0.05	1.36	14, 718	0.17 (nss)	0.05	1.71	0.05	0.03	0.82
Year 3 (2020)	354	0.15	4.09	14, 692	<.001***	0.05	1.71	0.15	0.08	0.99

* Number of groups and dependent variables: 3/7.

coefficients are slightly less stringent than conventional psychometric scales due to the increased complexities presented by portfolios in comparison to other literacy performances, such as essays generated under timed conditions.

Post-Hoc Power Analysis

Post-hoc power analyses (see Table 9) were conducted for the three assessments to test the null hypothesis, which assumes there is no difference between the three linguistic cohorts. Post-hoc power analysis can determine the statistical strength of the probability that a model registers hypothetical differences among populations in an assessment sample. In controlled studies, some argue that post-hoc power analysis is unnecessary when applied to already statistically significant results, adding little to explanations of non-statistically significant results (Heckman et al., 2022). However, in exploratory empirical studies such as this one, post-hoc power analysis can add

nuance to statistically significant results. The three cohorts—MoE, MuE, Mu—are independent variables, and the scores for each rubric category are dependent variables. The post-hoc power tests used a MANOVA: Global Effects F-test (Faul et al., 2007) to determine whether differences in effect sizes and power coefficients might complicate inferences drawn from the assessment data.

Disparate Impact Analysis

For each ePortfolio assessment, disparate impact analysis was performed using Stata software. Each analysis used mixed effects linear regression to model the rating of each rubric area on the interaction with race/ethnicity and then with primary home language variable, low income, and first generation. The scores were averaged across raters and treated as continuous numerical values. No corrections were made for multiple comparisons. P-values are statistically significant at alpha < 0.05. For each analysis, the respective reference categories were RaceEth=Asian, Multilingual English (MuE), First Generation=0 (no), and Low Income=0 (no). Each of the other categories were compared to the reference categories. Data on race/ethnicity outside of Asian, Hispanic, and White, Non-Hispanic were excluded due to small numbers. For each DI analysis, assessment results (see Table 3 in Supplemental Materials) were merged with low-income and first-generation data, which led to reduced sample sizes (see Table 4 in Supplemental Materials). The disparate impact method used here relies on previous writing studies research (Kelly-Riley & Elliot, 2016; Poe et al., 2014; Poe & Cogan, 2016), which describes "disparate impact analysis as a validation tool for understanding the local effects of writing assessment on diverse groups of students" (Poe et al., 2014, p. 288).

Research Setting

In Fall 2013, U.C. Irvine's composition program introduced a final ePortfolio into the assessment ecology of its general education capstone course. In response to changing demographics with the increase of undergraduate international students, a trend across many postsecondary institutions in the United States (Hussar & Bailey, 2013; Queen, 2017), curricular adaptations were needed as linguistic heterogeneity at an already polyglot campus deepened. With upwards of 75% of incoming students transferring well-cultivated multilingual resources, ePortfolios with metacognitive focal points were conceived to give teachers more information about student responses to the distinctive assessment ecologies of their courses. Equity-minded teachers developed new insights about formative and summative assessments as current-traditional assumptions (Young, 1978) about literate proficiencies became an active sedimentary layer within a progressively evolving curricular ecology in which ePortfolios instantiated reflective and evidence-based practices for both students and teachers (CCCC, 2015; Yancey, 2009).

The samples for the three ePortfolio assessments were collected as a matter of programmatic practice from Spring 2016 to Spring 2020 across regular-year academic quarters from sections of the Program's general education capstone class. This research-writing course—with social justice advocacy and information literacy focal points—delivers approximately 250 sections with distinctive themes to approximately 6,000 students annually. As students deepen their knowledge of exigent sociopolitical problems offered by the broad themes of their sections—among them mass incarceration, climate change, medical humanities, education, and constitutional law—they assemble final portfolios by engaging reflective prompts and gathering evidence of their learning.

A common final ePortfolio prompt establishes the task's contours for all sections and asks students to document and offer balanced assessments of their learning across the term by delivering compelling and analytically incisive claims in a multimodal reflective introduction that curates the substantive arrangement of the portfolio's sections and artifacts. This assignment prompt establishes a common method for the summative grading of ePortfolios across classes and offers students guiding questions for metacognitive analysis and suggestions for organizing the sections of their portfolios framed by emphases on knowledge transfer, composing and invention strategies, revision and crafting insights, researching and information literacy insights, and other elements of their choosing, such as the selection of meaningful artifacts from other courses and reflective narration situated within their personal histories.

The three assessments took place in the summers of 2017, 2018, and 2020 and used a rubric whose form and method derive from the common ePortfolio assignment prompt utilized during the academic year. These exercises in formative programmatic assessment attune teachers to the common ePortfolio instrument. As complexly contextualized literacy performances, the ePortfolios that make up the assessment samples bridge classroom activity systems and the larger programmatic assessment ecology of which they are a part, thereby establishing the naturalistic conditions for curricular attunement and validation analysis.

After several years of adapting pedagogically to ePortfolios and assessing them, and as the Program honed approaches to the metacognitive aspects of the capstone course's curriculum, a thick longitudinal body of assessment data had accumulated that here serves as the foundation for an exploratory analysis of the ePortfolio instrument's fairness, or what the field of writing assessment would conceive as its validity.

Results

Inter-rater Reliability Analysis

Generally, and by the "whale scale," the ePortfolios were assessed reliably across the three assessments and all rubric areas (see Table 3 in Supplemental Materials). Coefficients for both consensus and consistency estimates range from low medium to the high medium range across the first two assessments, with one and the same rubric trait—multimodality—registering high agreement coefficients. For the third assessment, the non-adjudicated kappa coefficients in all rubric categories register high levels of agreement and document a pattern of upward movement in reliabilities for all rubric areas.

The inter-rater reliability statistics reported for non-adjudicated portfolio scoring suggest solid to strong consensus judgments attended to by high levels of consistency across mean scores for all rubric categories. Notably, both the kappa and Spearman coefficients for every rubric trait outside of the narration trait on the first assessment register stronger estimates than the conventions and mechanics area. Read reliably according to these estimates, all of which are supported by statistical significance indicators at <.001, the longitudinal scoring of ePortfolios suggests that teachers as assessors have seemingly developed common dispositions related to the comprehension of rhetorical metacognition's traits as important indicators of academic performances tied to specific learning objectives.

ePortfolio Assessment Results

For the first assessment (2017), students who have been multilingual throughout their lives—Multilingual English (MuE)—attain the highest scores for all rubric categories except for the conventions and mechanics trait in which Monolingual English (MoE) students rate first. Multilingual (MU) students take third in this area that defines markers of traditional literate proficiency within educational institutions. Notably, readers are trained to assess this trait by its conventional contours tied to editing, grammar, usage, and sentence-level characteristics, and by students' awareness of the rhetorical effects of this trait and its sub-elements. But the Mu students—the cohort with the largest number of international students—place second and MoE students third in ratings for the first holistic impression, the argumentation trait, arrangement, and multimodality. All results are statistically significant except for multimodality.

The second assessment, undertaken in Summer 2018, shows no statistically significant differences in the comparison of scores for each rubric question, except for conventions and mechanics, which meets the threshold of <.05. Scoring in this area breaks down along traditional demographic lines, with MoE first, MuE coming next, and then Mu. The other results document key similarities to the first assessment with MuE students scoring the highest in four of the seven rubric areas.

Undertaken in Summer 2020, the third assessment includes student work from the first full term of the COVID-19 pandemic. This sample draws from two quarters delivered in classrooms and the first quarter of remote instruction, a moment of high stress. Several interesting changes emerge. The first is the development of higher reliability estimates across all rubric categories, and scores across all areas are statistically significant at <.01 and <.001 (see Table 3 in Supplemental Materials). With this assessment, unlike the other two, the MoE cohort places first in all areas, except for narration where MuE is highest, but differences between the MoE and the MuE cohort are very close and equal in the argumentation area. All three cohorts perform well here, but the Mu cohort tracks lower in comparison than it had with the previous assessments, and this cohort is again third in conventions and mechanics.

Post-hoc Power Analysis

The results from the three post-hoc power analyses show that all three assessment samples meet the accepted minimum threshold of 0.80 for generalization, suggesting that each can detect effect size differences among the cohorts being compared. By conventional scaling (Cohen, 1988), effect size indicators range from the upper end of the medium tier for assessment 1, to middle of the small tier for assessment 2, to the medium tier for assessment 3. The resulting power registers for the first and third assessments come in at 0.99 and the second at 0.82.

Disparate Impact Analysis

The comparisons of assessment results by race/ethnicity (see Tables 4, 5, 6, and 7 in Supplemental Materials)—broad demographic cohorts of defined as Asian, Hispanic, and White, Non-Hispanic—reveal no statistically significant results across the three assessments for all rubric areas, except for the conventions and mechanics trait for just the first assessment in which White, Non-Hispanic and Hispanic students score higher than Asian students, with Hispanic students scoring the highest. With cohort comparisons defined solely by race/ethnicity, a large proportion

of the Mu cohort, which is composed predominantly of international students, folds into the broader Asian population.

When moving to three-way intersectional comparisons (see Tables 5, 6, and 7 in the Supplemental Materials), a more nuanced statistical picture emerges. Within the MuE cohort for the first assessment, the intra-MuE grouping that is neither first generation nor low income documents the highest means within this group across all rubric categories. Only one comparison is statistically significant, in conventions and mechanics, for MuE students who are both first generation and low income (see Table 5 in Supplemental Materials). When comparing the MuE base cohort that is neither first gen nor low income across the MoE and Mu cohorts and their permutations, several instances of statistically significant differences can be seen. Monolingual English students (MoE) who are both first gen and low income score lower in statistically significant ways than the Multilingual English (MuE) base cohort that is neither first gen nor low income in all rubric areas except arrangement and narration. These data document discernable patterns suggesting that monolingual English students who are both first generation and low-income perform lower than multilingual English students who are neither. Additionally, the MuE cohort with neither first generation nor low-income markers manifests more developed dispositions of metacognition across all three cohorts and their permutations in all rubric areas, except when compared with the Multilingual (Mu) cohort that is low income but not first generation in the areas of the first holistic variable, argumentation, arrangement, and narration.

With the second assessment, nearly every indicator of statistical significance across rubric categories and the three cohorts and their first gen/low-income permutations goes away, outside of two statistically significant indicators within the MoE cohort that is first generation and not low income in which the multimodality and the second holistic variables score lower than the MuE base cohort. Moreover, first generation and low-income students within each cohort score higher with some frequency than the MuE No/No cohort (see Table 6 in Supplemental Materials). These prevailing trends hold sway with the disparate impact analysis for the third assessment (see Table 7 in Supplemental Materials). There are eight indicators of statistical significance, all at <.05, seven of which document first-gen and/or low-income cohorts within the MoE and MU cohorts performing slightly better than the MuE that has neither marker. Across these intersectional demographic comparisons from all three assessments, distinctions among cohorts do not consistently document prevailing patterns in which first generation and/or low-income students across language backgrounds perform lower than students who have neither marker. Moreover, the results do not document monolingual English students outscoring the other two language cohorts systematically.

Discussion

To respond to this study's guiding question—Can ePortfolios be instruments of fairness in a local assessment ecology?—a return to the four sources of validity evidence (Kane, 2006; Poe et al., 2014) is warranted. In terms of scoring evidence, data were derived from a thick longitudinal body of ePortfolio assessment data as sourced from the three ePortfolio assessments. Two forms of inter-rater reliability estimates support inferences drawn from the ePortfolio assessment results, and adjudication was not performed to address discrepancies in scoring. With the resulting reliability coefficients offering conventionally acceptable reliability levels, it can be reasonably deduced that the reliability estimates also serve as generalization evidence in that they register the importance

of knowledge accumulated through the applied act of teaching within the same curricular ecology and thereby establish a first validity inference: that readers reliably assessed ePortfolios from within an common assessment ecology of information about rhetorical metacognition, the multifaced literacy construct described by the assessment rubric.

In their study of fairness as an integrative principle for writing assessment, Kelly-Riley et al. (2016) "hold that reliability information is an important prerequisite to evidence of validity and fairness" (p. 102). Following this method, and with the reliabilities for the assessment data strongly supported by rigorous statistical analysis, a second inference can be deduced from further scoring data. Assessment evidence, generated by the rigorous statistical methods deployed, suggests that organizing assessment cohorts by language difference—as defined by the primary home language variable whose validity is attained in the process—creates the quantitative foundation for validation analysis of the metacognitive assessment instrument. These deductions set up several subsequent inferences.

While reliability estimates for both consensus and consistency measures across all rubric categories fall within the medium strength area for the first assessment, the multimodality trait has the second highest reliability rating to the second holistic impression. When considering that multimodality is the only trait that does not register statistical significance in the comparison of mean scores, this data point suggests that this aspect of rhetorical metacognition, which is defined essentially by awareness of the rhetorical effects of hybrid visual/textual arrangements, is not recognized by readers as suggestive of difference between the linguistic cohorts. When this insight is combined with the conventions and mechanics score on the first assessment, which documents score differences at the highest level of statistical significance with the Mu cohort scoring lowest, a major inference emerges. Together, these data suggest that raters are aware of and reliably assess current-traditional expectations for literacy performances defined by conventions of standard edited academic English while they simultaneously recognize and reliably assess multimodality as a rhetorical principle of arrangement. Data from the second assessment supports this deduction. Notably, scoring differences on the holistic questions are not statistically significant, and when considering that the differences in the mechanics and conventions area are statistically significant and that the MuE students score below the monolingual English students in this area, additional validation inferences can be drawn from the weighty indicators across the first two assessments, in which the MuE cohort receives the highest ratings in most categories: (a) assessors recognize traditional indicators of literate proficiency, even as they capably evaluate and value rhetorical metacognition as a register of literate proficiency across linguistically heterogeneous cohorts; (b) assessment data suggests that assessors recognize multiple and distinctive literacies that are enabled by the metacognitive ePortfolio assignment.

The scoring data from the third assessment complicates the inferences derived from the first and the second. This assessment includes samples from the first two quarters of the COVID-19 quarantine, which were not treated separately or systematically in the sampling method. The third assessment shows the MoE cohort scoring the highest generally with the MuE cohort following closely behind and then the Mu cohort third in statistically significant results. This alignment in scoring, which might be expected in an assessment system privileging English monolingualism, suggests that more generalization evidence and extrapolation evidence are needed to further examine the representative nature of the cohorts, the validity of the linguistic variable, and to seek more fine-grained indicators of intersectional difference in this assessment ecology.

As generalization evidence, the three power analyses in Table 9 suggest that the primary home language variable can explain mean differences in ePortfolio scores. While further validating the use of this variable for portfolio assessment, these power analyses substantiate the preceding inferences, themselves derived from the weighty body of longitudinal assessment data. Nevertheless, when considering the differences in the effect sizes, that smaller effect sizes in this analysis may suggest a leveling effect across the cohorts and assessments—perhaps an indicator of fairness—the lower power percentage for the second assessment alongside the varying levels of statistical significance found across the ratings for the second assessment and when combined with the scoring alignment of the third assessment suggest that a stratified sampling method is warranted to offer further support to or to undermine claims about the validity of the linguistic variable.

As extrapolation evidence, disparate impact analysis (see Tables 4, 5, 6, and 7 in Supplemental Materials) uses the aforementioned scoring and generalization evidence to test assessment outcomes for statistical significance measured first by comparing demographic categories of race/ethnicity and then by comparing the linguistic cohorts within themselves and across them by considering combinations of intersectional demographics. With the race/ethnicity comparisons demonstrating an absence of statistical significance across the three assessments in all categories except for the conventions and mechanics trait from the first assessment, this data supports the inference that broad racial typologies may not provide meaningful statistical distinctions in this assessment ecology. Moreover, even when viewed through the lens of the language variable as it enregisters intersectional demographics, the absence of meaningful and consistent patterns of statistical significance indicators is notable in this ecology when assessing for rhetorical metacognition.

Conclusion

To conclude, the four types of validity evidence presented by this formative longitudinal study suggest together that a principle of fairness in assessing ePortfolios may be salient within the local assessment ecology studied. From three ePortfolio assessments, whose samples are defined by an institutionally steeped linguistic variable, which is used to assess a complex writing construct defined as rhetorical metacognition, a dense body of validation evidence and attendant inferences emerge to validate the guiding linguistic variable as a telling indicator of language difference that enregisters important intersectional demographics. Moreover, from the weighty compilation of longitudinal data and inferences, an encompassing inference materializes to inform a local theory of fairness and serves as consequential validity evidence: that teacher-assessors recognize traditional indicators of literate proficiency but credibly recognize and assess metacognition as an indicator of rhetorical literacy. When considering the synthesis of data and evidence generated by the quantitative methods used here, it would seem, to return to the definition of fairness in writing assessment forwarded by Elliot (2016), that ePortfolio pedagogy and its metacognitive writing construct create accessible opportunities for students across a deeply multilingual and racially heterogeneous population to perform their acquisition of a foundational aspect of post-secondary literacy by drawing from antecedent knowledge and idiosyncratic resources. The institutionally defined contours of the ePortfolio assessment instrument seem to enable the least advantaged, perhaps suggesting that a principle of fairness in this assessment ecology may be apparent. Nevertheless, the critical study of fairness through rigorous validation analysis must be ongoing

as an ethical principle of assessment design and curricular attunement in a continuous search for social justice in a local writing assessment ecology.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. https://www.aera.net/publications/books/standards-for-educational-psychological-testing-2014-edition

Bawarshi, A. (2017). Economies of knowledge transfer and the use-value of first-year composition. In B. Horner, B. Nordquist, & S. Ryan (Eds.), *Economies of writing: Revaluations in rhetoric and composition* (pp. 87–98). Utah State University Press; University Press of Colorado.

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

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

Black, L., Daiker, D., Sommers, J., & Stygall, G. (Eds.). (1994). *New directions in portfolio assessment: Reflective practice, critical theory, and large-scale scoring*. Heinemann/Boynton-Cook.

Bowles, S., & Gintis, H. (2002). Schooling in capitalist America revisited. *Sociology of Education, 75*(1), 1–18.

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

Bryant, L. H., & Chittum, J. R. (2013). ePortfolio effectiveness: A (ill-fated) search for empirical support. *International Journal of ePortfolio, 3*(2), 189–198.

Canagarajah, A. S. (2013). Negotiating translingual literacy: An enactment. *Research in the Teaching of English, 48*(1), 40–67.

Che, C. (2022). Mind the (linguistic) gap: On "flagging" ESL students at Queensborough Community College. In J. Nastal, M. Poe, & C. Toth (Eds.), *Writing placement in two-year colleges: The pursuit of equity in postsecondary education* (pp. 191–222). The WAC Clearinghouse; University Press of Colorado.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Conference on College Composition and Communication [CCCC]. (2015). *Principles and practices in electronic portfolios*. https://cccc.ncte.org/cccc/resources/positions/electronicportfolios

Conference on College Composition and Communication [CCCC]. (1974). Students' right to their own language (SRTOL). *College Composition and Communication, 25*(3), 1–18.

Council of Writing Program Administrators (CWPA), National Council of Teachers of English (NCTE), and National Writing Project (NWP). (2011). *Framework for success in postsecondary writing* [White Paper]. https://wpacouncil.org/aws/CWPA/asset_manager/get_file/350201?ver=7548

Cushman, E. (2016). Decolonizing validity. *Journal of Writing Assessment, 9*(1). https://escholarship.org/uc/item/0xh7v6fb

Davila, B. (2017). Standard English and colorblindness in composition studies: Rhetorical constructions of racial and linguistic neutrality. *WPA: Writing Program Administration, 40*(2), 154–173.

Driscoll, D. L., & Zhang, J. (2022). Mapping long-term writing experiences: Operationalizing the writing development model for the study of persons, processes, contexts, and time. *Composition Forum, 48*. https://compositionforum.com/issue/48/mapping.php

Elliot, N. (2015). Validation: The pursuit. *College Composition and Communication, 66*(4), 668–687.

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

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Georghiades, P. (2004). From the general to the situated: three decades of metacognition. *International Journal of Science Education, 26*(3), 365–383.

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.

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*, 355–392.

Gorzelsky, G., Driscoll, D. L., Paszek, J., Jones, E., & Hayes, C. (2017). Cultivating constructive metacognition: A new taxonomy for writing studies. In C. M. Anson & J. L. Moore (Eds.), *Critical transitions: Writing and the question of transfer* (pp. 215–246). The WAC Clearinghouse; University Press of Colorado.

Hamp-Lyons, L., & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory, and research*. Hampton Press.

Haswell, R., & Elliot, N. (2019). *Early holistic scoring of writing: A theory, a history, a reflection*. Utah State University Press.

Heckman, M. G., Davis, J. M., & Crowson, C. S. (2022). Post hoc power calculations: An inappropriate method for interpreting the findings of a research study. *Journal of Rheumatology, 49*(8), 867–870. https://doi.org/10.3899/jrheum.211115

Horner, B., & Trimbur, J. (2002). English only and U.S. college composition. *College Composition and Communication, 53*(4), 594–630.

Horner, B., NeCamp, S., & Donohue, C. (2011a). Toward a multilingual composition scholarship: From English-only to a translingual norm. *College Composition and Communication, 63*(2), 269–300.

Horner, B., Lu, M. L., Royster, J. J., & Trimbur, J. (2011b). Language difference in writing: Toward a translingual approach. *College English, 73*(3), 303–321.

Hussar, W. J., & Bailey, T. M. (2013). *Projections of education statistics to 2022*. National Center for Education Statistics, U.S. Department of Education. https://nces.ed.gov/pubs2014/2014051.pdf

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.

Inoue, A. B. (2009). The technology of writing assessment and racial validity. In C. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 97–120). IGI Global.

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

Kelly-Riley, D., & Elliot, N. (2021). *Improving outcomes: Disciplinary writing, local assessment, and the aim of fairness*. Modern Language Association.

Kelly-Riley, D. (2011). Validity inquiry of race and shared evaluation practices in a large-scale, university-wide writing portfolio assessment. *Journal of Writing Assessment, 4*(1). https://escholarship.org/uc/item/7m18h956

Kelly-Riley, D., & Whithaus, C. (2016). Introduction to the special issue on a theory of ethics for writing assessment. *Journal of Writing Assessment, 9*(1). https://escholarship.org/uc/item/8nq5w3t0

Kelly-Riley, D., Elliot, N., & Rudniy, A. (2016). An empirical framework for ePortfolio assessment. *International Journal of ePortfolio, 6*(2), 95–116.

Lam, R. (2017). Taking stock of portfolio assessment scholarship: From research to practice. *Assessing Writing, 31*, 84–97.

Lederman, J. (2023). Validity and racial justice in educational assessment. *Applied Measurement in Education, 36*(3), 242–254. https://doi.org/10.1080/08957347.2023.2214654

Lee, J. W. (2018). *The politics of translingualism: After Englishes*. Routledge.

Lu, M., & Horner, B. (2013). Translingual literacy, language difference, and matters of agency. *College English, 75*(6), 582–607.

Messick, S. (1989). Validity. In R. L. Lin (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). American Council on Education; Macmillan.

Matsuda, P. K. (2006). The myth of linguistic homogeneity in U.S. college composition. *College English*, 68(6), 637–651.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5–12.

Negretti, R. (2012). Metacognition in student academic writing: A longitudinal study of metacognitive awareness and its relation to task perception, self-regulation, and evaluation of performance. *Written Communication, 29*(2), 142–179.

Negretti, R., & Kuteeva, M. (2011). Fostering metacognitive genre awareness in L2 academic reading and writing: A case study of pre-service English teachers. *Journal of Second Language Writing, 20*(2), 95–110.

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

Perryman-Clark, S., Kirkland, D. E., & Jackson, A. (Eds.). (2014). *Students' right to their own language: A critical sourcebook*. Bedford/St. Martin's.

Poe, M., Oliveri, M. A., & Elliot, N. (2023). The *Standards* will never be enough: A racial justice extension. *Applied Measurement in Education*, 36(3), 193–215. https://doi.org/10.1080/08957347.2023.2214656

Poe, M., & Elliot, N. (2019). Evidence of fairness: Twenty-five years of research in *Assessing Writing*. *Assessing Writing, 42*, 100418.

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.

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

Poe, M., & Cogan, J. A. (2016). Civil rights and writing assessment: Using the disparate impact approach as a fairness methodology to determine social impact. *Journal of Writing Assessment, 9*(1). https://escholarship.org/uc/item/08f1c307

Poe, M., Elliot, N., Cogan, J. A., & Nurudeen, T. G. (2014). The legal and the local: Using disparate impact analysis to understand the consequences of writing assessment. *College Composition and Communication, 64*(4), 596–597.

Queen, B. (2017). Class size for a multilingual mainstream: Empirical explorations. *WPA: Writing Program Administration, 40*(2), 98–128.

Randall, J., Slomp, D., Poe, M., Oliveri, M.A. (2022). Disruptive White supremacy in Assessment: Toward a justice-oriented, anti-racist validity framework. *Educational Assessment, 27*(2), 170–178. https://doi.org/10.1080/10627197.2022.2042682

Randall, J., Poe, M., Oliveri, M. A., & Slomp, D. (2024a). Our validity looks like justice. Does yours? *Language Testing, 41*(1), 203–219. https://doi.org/10.1177/02655322231202947

Randall, J., Poe M., Oliveri, M.A., Slomp, D. (2024b). Justice-oriented, antiracist validation: Continuing to disrupt White supremacy in assessment practices. *Educational Assessment, 29*(1), 1–20. https://doi.org/10.1080/10627197.2023.2285047

Raudenbush, S., & Kasim, R. (1998). Cognitive skill and economic inequality: Findings from the national adult literacy survey. *Harvard Educational Review, 68*(1), 33–80.

Rhodes, C. R., Clonan-Roy, K., & Wortham, S. E. F. (2020). Making language 'academic': Language ideologies, enregisterment, and ontogenesis. *Language and Education, 35*(6), 522–538. https://doi.org/10.1080/09500782.2020.1797771

Rosa, J. (2019). *Looking like a language, sounding like a race: Raciolinguistic ideologies and the learning of Latinidad*. Oxford University Press. https://doi.org/10.1093/oso/9780190634728.001.0001

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist, 24*(2), 113–142.

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 278–298). Routledge.

Scott, B. M., & Levy, M. G. (2013). Metacognition: Examining the components of a fuzzy concept. *Educational Research and Evaluation, 2*(2), 120–131.

Stemler, S. E. (2019). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). https://doi.org/10.7275/96jp-xz07

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

Taczak, K., & Robertson, L. (2017). Metacognition and the reflective writing practitioner: An integrated knowledge approach. In P. Portanova, J. M. Rifenburg, & D. Roen (Eds.), *Contemporary perspectives on cognition and writing* (pp. 211–229). The WAC Clearinghouse; University Press of Colorado.

Urciuoli, B. (2009). Talking/not talking about race: the enregisterment of culture in higher education discourse. *Linguistic Anthropology, 19*(1), 21–39. https://doi.org/10.1111/j.1548-1395.2009.01017.x

Wardle, E., & Roozen, K. (2012). Addressing the complexity of writing development: Toward an ecological model of assessment. *Assessing Writing, 17*(2), 106–119.

Wardle, E. (2009). 'Mutt genres' and the goal of FYC: Can we help students write the genres of the university? *College Composition and Communication, 60*(4), 765–789.

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: The assessment of writing programs*. Utah State University Press; University Press of Colorado.

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

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50*(3), 483–503.

Yancey, K. B. (2009). Portfolios, circulation, ecology, and the development of literacy. In D. N. DeVoss, H. A. McKee, & R. Selfe (Eds.), *Technological ecologies & sustainability*. Computers and Composition Digital Press; Utah State University Press. https://ccdigitalpress.org/book/tes/05_yancey.pdf

Yancey, K. B., & Selfe, R. (1996). Portfolios, electronic, and the links between. *Computers and Composition, 13*(2), 135–146.

Yancey, K. B., Robertson, L., & Taczak, K. (2014). *Writing across contexts: Transfer, composition, and sites of writing*. Utah State University Press; University Press of Colorado.

Young, R. (1978). Paradigms and problems: Needed research in rhetorical invention. In C. R. Cooper & L. Odell (Eds.), *Research on composing: Points of departure* (pp. 29–48). National Council of Teachers of English.