RETROSPECTIVE.

A REFLECTIVE ANALYSIS: TOWARD FAIRNESS

Mya Poe
Northeastern University

What would test fairness bring to individual students? In many ways, this question is behind issues of fairness as debated in many other assessment journals for the last several decades. During the first two decades of *The Journal of Writing Assessment's* history, however, the question of fairness and justice for individual students has been shaped by a deep disciplinary commitment to the lived realities of writing assessment and wrestling with whether measurement theory helps us understand those realities.

Beyond *JWA*, researchers from different disciplinary contexts have long debated origin stories, developments in evidence-gathering, and implications for stakeholders. On one hand, educational measurement scholars have deliberated the expansive connotations of the term fairness (Boyer, 2020; Dorans & Cook, 2016; Gipps & Stobart, 2010; AERA, APA, & NCME, 2014). On the other hand, they have fiercely proscribed narrow administrations of fairness to test design and development, test administration, scoring, and score interpretation (Dorans & Cook, 2016). Fairness and bias reviews suggest that fairness is something that can be observed in textual analysis (ETS, 2014, 2016) or ferreted out in the analytic tools created by designers, ranging from general linear models developed by Cleary (1968) and in more advanced forms, in use today. Following the *Standards for Educational and Psychological Testing,* many measurement researchers conjoin fairness with validity:

> Fairness is a fundamental validity issue and requires attention throughout all stages of test development and use. . . . [F]airness and the assessment of individuals from specific subgroups of test takers, such as individuals with disabilities and individuals with diverse linguistic and cultural backgrounds . . . is an overriding, foundational concern, and that common principles apply in responding to test-taker characteristics that could interfere with the validity of test score interpretation. (AERA, APA, & NCME, 2014, pp. 49-50)

DOI: https://doi.org/10.37514/PER-B.2024.2326.2.24

What makes a test fair is, as Xi (2010) argues, "comparable validity for all *relevant* groups" (p. 147). Likewise, Zieky (2016) claims that "the fairness argument is an extension of the validity argument. The goal of the fairness argument is to present evidence that the test is fair for various groups within the test-taking population" (p. 96).

Researchers from the writing studies community, such as my colleagues Oliveri, Elliot, and I (2023) have argued that the *Standards* offer other affordances to increase fairness: (a) accessibility (unobstructed opportunity for diverse groups to have equal opportunity to take a test and demonstrate construct standing); (b) universal design (designing a test and its associated delivery environment to maximize usability by all test takers); and (c) opportunity to learn content that is culturally sustaining to their own communities (the degree to which test results need to be evaluated for maximum community impact). We believe that, "making good decisions about our writing assessment practices for all students means attending to the various ways that we understand the impact of assessment on our students" (Poe & Cogan, 2016, p. 605). At the end of the day, no test is culture free, and assessment is about its effects on diverse individuals and communities. As I have argued elsewhere (Poe & Cogan, 2016), the authors of the *Standards* left the larger challenge for fairness—i.e., the relationship of assessment to social consequences—relatively untouched.

In each of the articles in this section of *Considering Fairness and Aspiring to Justice*, the authors wrestle with what disciplinary theories and methods should we use to "form attitudes or induce actions in other human agents" (Burke, 1950, p. 41). But to move directly to the articles themselves does not seem exactly right. I read each of the chapters in this section of *Considering Fairness and Aspiring to Justice*—exciting work by Peggy O'Neill, Bob Broad and Michael Boyd, Asao B. Inoue, and David Slomp—with Brad, my brother, born in September of 1966, in mind. Yes, I want to address what the author's historical situatedness means for the way they conceive of fairness. And I want to address what kinds of social implications each author considers. But, first, I want to talk about Brad.

FRAMING FAIRNESS

According to my mother, Brad was a colicky infant. He rarely cried and showed little emotion as a toddler but took much interest in mechanical objects and family pets. When he was older, he built elaborate train tables with lights and gates that were operated by electrical circuits he had soldered. He would later spend hours reading books backwards and forwards, often selecting massive books on technical subjects as well as *Mad* magazine. He was disorganized, his

handwriting was a scrawl, and his sense of sci-fi fan humor was often described as "warped" by my parents.

In school, he made few friends and seemed disinterested in schoolwork. In kindergarten school, Brad was diagnosed with a moderate sensorimotor development delay. My mother who was a teacher tried to coax Brad along, encouraging and working with him on balance and coordination. She tried to help him show emotion, which would only result in outbursts of anger. Later, IQ test results were very high, yet Brad often earned average and below-average grades. My father, who had dropped out of high school at age 17 often reacted with rage, unable to understand why someone so "smart" could be so "lazy." No amount of yelling and badgering and humiliation motivated Brad. Brad simply fell silent, seemingly emotionally vacant.

Brad struggled through high school but was able to enroll in college. He couldn't get into the engineering program he wanted, so he became a business major. As an undergraduate business major, he commuted to college, splitting his time between a job at an auto parts store, helping my father on the farm, and squeezing in homework. He could not manage the long commute, the demands of my father and the job, and the demands of college. He graduated college with barely a C average yet got a probationary enrollment in an evening MBA program at the same college. In graduate school, Brad struggled, once again trying to manage a life split across worlds and avoiding the required group projects of an MBA program. Yet, there was one thing that graduate school brought Brad—the VAX machine, an early supercomputer the size of a small refrigerator. Brad spent hours at the computer bank.

To this day, no one in my family knows what he was coding because Brad hung himself in 1993 at the age of 26. Unlike many people who leave textual artifacts of their lives behind—notes, scribbles of random ideas, tickets, receipts, documentation—Brad's life was undocumented except for some banking documents, some school notebooks, car manuals, and a letter from the MBA program stating that Brad was going to be expelled for poor grades. I saved a notebook from his desk—a notebook from his MBA studies—and the two exams that were tucked inside. Pages of his notes from his college notebook are illegible. Some are half-written. Others are filled with technical terms and graphs with no meta-commentary about the content. Much of the notebook is empty. A paper from a management course on leadership showed a grade of 19 out of 20 points. A fall 1992 final exam from an accounting class in the MBA program showed that he received 107.5 out of 120 points (89.6%). One written question asked test-takers to select a regression model and provide specific reasons for the selection of that model for a fictional character named "Alf" (perhaps a nod to the *Mad Magazine* character). The professor noted some comments in red on

Brad's written response, including -3 points noted next to a postscript that Brad had written: "Alf went on to run the Dan Quayle Presidential campaign in '96, when Dan was decisively defeated by Alfred E. Newman."

I start with this story about my brother Brad because Brad's story illustrates how what we see through assessment is deeply shaped by historical context**.** Brad died before the "autism epidemic" (Nuwer, 2016), but it is likely that if my brother was born today, he would be diagnosed on the autism spectrum (Hannant, 2016; Hyman et al., 2020). In the 1960s and 1970s, assessment instruments like sensory and motor development tests were used to determine physical delays that might indicate cognitive delays. IQ tests were common in schools like the one Brad attended to track students. Classroom assessment technologies, especially in the disciplines, were still largely summative, and notions of "progress" through degree programs were still largely tied to course grades. These systems of assessment accumulated to provide a measurement-based narrative of a child.

Brad's story also reminds us of the social implications of assessment. In the 1960s and 1970, if such assessment technologies existed to "measure" autism, it is unlikely that Midwestern suburban lower-middleclass schools would have had such assessment technologies to understand children like Brad. What they did have were assessment instruments like motor development tests and intelligence tests that had been refined into codified instruments delivered through school volunteers and classroom teachers. Furthermore, my father was unable to reconcile tests that showed competing narratives of his child—one delayed and one gifted. For my father, tests carried enormous social prestige. They were scientific diagnostic instruments that told the truth about his child. For people like my high school drop-out father, whose father and mother had eighth and sixth grade educations respectively, the message of tests was absolute. Concepts like "intelligence" were highly valued because Appalachian whites work within a cultural context in which they are often regarded as exoticized isolates yet also portrayed as inbred, immoral, and stupid. For my father, to have his son be labeled "gifted" was scientific proof that he *personally* was not genetically inferior. For my father, there was a familial obligation to live up to the term "gifted." No test designer was in the room when my father humiliated and kicked my "gifted" brother for getting bad grades, claiming that poor grades and test scores were merely the result of being "lazy."

My understanding of educational measurement has opened doors to understanding my brother's life left in assessment artifacts. The assessment artifacts of Brad's life provide a consideration of fairness and the impacts of assessment on different humans. Here was a student for whom assessment provided a narrative about his purported inner potential and documented his outward failings and blamed him for those failings. Here is a student who ultimately graduated from college and was enrolled in a graduate program when he died. By one

benchmark—college completion rates—Brad was a success and soon-to-be-failure when he died.

What would test fairness have brought Brad? Likely little if we were to rely on measurement theory as a guide. Maybe more if we rely on articles such as those to which I now turn.

PAST AS PROLOGUE

During most of the twentieth century, writing assessment researchers have had a love/hate relationship with the field of measurement. On one hand, researchers and teachers have long fought the over-reach of the testing industry into writing classrooms and programs. On the other hand, we have been exhorted by scholars within writing studies to adopt measurement theories related to validity, reliability, and fairness to improve the design and use of writing assessment. Those exhortations were strongest in the late 1990s into the early 2000s, yet today we see those theoretical connections –-citational pathworks—happening between measurement and writing studies. In these citational pathways, we can trace how researchers within any historical context have certain vantage points from which they see the social implications of assessment—i.e., the ways in which assessments are being used, the targets of assessment, and the ways assessment is connected to other institutional and social systems.

In documenting the work of William L. Smith at the University of Pittsburgh during the 1980s and 1990s, Peggy O'Neill situates Smith's work within "the larger context of educational measurement theories, placement testing, and holistic scoring" and argues that Smith's work is "an example of how systematic, ongoing validity inquiry can not only lead to better—more valid—local assessment but also contribute to the larger field of writing assessment" (p. 34). For everyone who has read his work, it is clear that Smith was an innovator; O'Neill saw that innovation and aimed to advance commonality. To make the case for the value of validity inquiry, O'Neill describes Smith's embrace of measurement theory:

> According to Smith (1998), there is a "paucity of validation research" (p. 3) in writing assessment, which stems from several different but interrelated problems: a lack of understanding of key concepts such as validity and reliability; an overemphasis on achieving reliability; a lack of understanding of what validation inquiry entails; and a failure to articulate the theoretical constructs underlying writing assessments. (p. 31)

O'Neill connects Smith's intellectual work with seminal measurement scholars Lee Cronbach, Samuel Messick, and Pamela Moss, especially in

terms of their work on moving the field toward an argument-based model of validity.

The punchline for O'Neill is that "validation arguments are rhetorical constructs that draw from all the available means of support" (p. 32). From this vantage—in fact, a prescient one that illustrated the importance of interpretation and use of arguments advanced a decade later by Kane (2013)—she then draws a connection to writing studies scholar Brian Huot's work on writing assessment and validity. This citational pathway between measurement and writing studies ultimately allows O'Neill to claim that "this [interdisciplinary] approach to writing assessment would support the processes and theories associated with literacy, leading to more theoretical alignment between actual literate practice and the assessment of it (p. 33). In short, measurement theory, O'Neill proposed, would allow writing assessment researchers to theoretically align the teaching and assessment of writing.

The use of measurement theory for alignment between assessment and teaching is certainly evidenced in Smith's approach to assessment research at the University of Pittsburgh. At the University of Pittsburgh he investigated the local ways that test decisions were being made. He believed that the initial data on misplacement via teachers' readings of student essays were erroneous. A singular or double reading of student writing and replacement rates were insufficient. Student impressions were important as were teachers' perceptions, especially their perceptions over time:

> teachers' perceptions of students change considerably across the course of the semester. If gathered too early in the semester, teachers don't have enough evidence on which to base their decision; if gathered too late, teacher perception correlates very highly with the students' final grades, indicating that the students' actual performance is evaluated, not their potential. Smith concluded that teacher perception data should be collected during Weeks 3 through 5 of a 15-week semester. (p. 40)

In studies of rater reliability, Smith also found that raters' decisions varied by many factors, including raters' teaching experience, the course the rater most recently taught, when raters knew they were being tested, when raters scored as split-resolvers, when raters "made decisions about students, instead of merely judging texts," and when raters could not match students to a specific course (p. 58). Smith anticipated the later work of Dryer and Peckham (2014) and their emphasis on adopting an ecological view of processes in which, down to the level of the tables at which raters sat, differences occurred.

When Smith turned his attention to student performance, he "found that there were significant factors that influenced their performance, but that he could not control for them" (p. 41). Rather than pursuing studies of factors that influenced student performance, Smith focused on the programmatic context of writing assessment. He developed the expert-reader model in which raters place students into classes based on prototypes.

Regarding fairness, Smith's approach brings a longitudinal perspective. He allowed for judgments to change over time as teachers learned more about students. Yet, because he could not "control" for the factors that influenced student performance, perhaps including such things as testing histories, cultural context, and emotional well-being, Smith chose to not pursue further investigation. Valuable as his work was, it operated within a measurement paradigm of replicability in which further work was suspended for fear of contaminating the validity argument. Purpose pluralism was yet to come and, ironically, it was to come from a UK measurement researcher calling for assessment designs that should leverage "a multiplicity of assessment purposes simultaneously" (Netwon, 2017, p. 5).

SITUATED ASSESSMENT

Published in 2005, Bob Broad and Michael Boyd's 2005 "Rhetorical Writing Assessment: The Practice and Theory of Complementarity" also focuses on innovation in the field of writing assessment, arguing that "writing portfolio assessment and communal (shared, dialogical) assessment are two of our field's most creative, courageous, and influential innovations" (p. 51). As is the case with O'Neill, Broad, and Boyd point to Huot to uncover the "'epistemological basis' . . . on which these new principles and procedures are built" (p. 54). In the twenty-first century, context would become everything.

Looking to advances in psychometrics—note that Broad and Boyd Note chose the term "psychometrics" in lieu of "measurement" as a way to emphasize the statistical quality of the research described by Pamela Moss and Lee Cronbach—they see promise in these changes akin to changes in classical physics and quantum physics in which Niels Bohr questioned the effect of "measuring instruments" on the phenomena being measured: "Quantum physics, in opposition to the classical version, accepts that ultimately all knowledge is indeterminate because the methods we use and the vantage points from which we obtain evidence substantially alters the evidence itself" (p. 55). Returning to measurement, they cite Egon Guba and Yvonne S. Lincoln's (1989) invocation of Bohr's complementarity principle as well as English Studies scholar Bernard Alford's dissertation in which he "draws on the work of quantum physicists Menas Kafatos and Robert Nadeau to focus our understanding of

the principle of complementarity" (p. 57). Broad and Boyd see Alford's work as a means to move "beyond objectivism and subjectivism" so that "we can verify postmodern claims to contingent truths through a process of bringing radically distinct constructs into dialogue with each other within established human communities" (p. 57).

It is from this citational path that Broad and Boyd argue, citing writing studies scholars James Berlin and Kathleen Yancey, that "the portfolio is a postmodern development" (p. 60) that "offered a way to move beyond grading of single pieces of writing to a process of 'collection, selection, reflection, and projection'" (p. 58). Communal writing assessment (CWA) they see as something even more radical: "The more radical shift is away from seeking and valuing homogeneity among judges to seeking and valuing diversity" (p. 68). In arguing for the potential of CWA, they note that CWA breaks from traditional notions of standardization in psychometrics.

Broad and Boyd refer to this epistemological change as a "velvet revolution in writing assessment" (p. 63) and argue that:

> [I]f we, the scholars and practitioners of writing instruction and writing assessment, hesitate further to develop and defend the epistemological base of these two practices, they will remain vulnerable to rear-guard actions by those still working within a positivist, a reactionary, or simply a budget-cutting framework. (p. 64)

Rear-guard action is a very real possibility, as Broad and Boyd caution that many measurement practitioners are reluctant to acknowledge such advances (p. 16). Ultimately, the theorization of writing assessment is a means to defend portfolios and CWA from "those wielding well developed and thoroughly institutionalized discourses such as those of positivist psychometrics" (p. 64).

In suggesting that assessment be about contradiction and multiplicity, Broad and Boyd point to a possibility "beyond the tired objectivist–subjectivist dichotomy" (p. 12). Communal writing assessment especially suggested the possibility of fairness with the multiplicity of readings and readers. Multiplicity in the ways that writing is assessed, however, does not extinguish power relations, invite understanding, or suggest pluriversal options. The difficulty of balancing community and multiplicity is nowhere more apparent than in Broad and Boyd's illusion to the Velvet Revolution. Like the Velvet Revolution in Czechoslovakia in 1989, whose reformers could not see the dissolution of the country four years later into two countries—the Czech Republic and the Slovak Republic—Broad and Boyd could not see that CWA would not become part of the mainstream discourse in writing studies. Yet, their belief in the value of CWA would give

rise to many newer forms of assessment and would include core values of CWA. There was not to be a single way forward—how could there be—when context was the key.

SOPHISTIC TURNS

Asao B. Inoue's (2007) "Articulating Sophistic Rhetoric as a Validity Heuristic for Writing Assessment" takes yet another approach in "bridg[ing] disciplines [of measurement and writing studies] by articulating validity in terms of rhetorical theory, and understanding ancient sophistic rhetorical positions as validity theory" (p. 67). Like O'Neill as well as Broad and Boyd, Inoue provides a citational chain through linkages of Cronbach, Messick, Moss, and Huot along with Lorrie Shepard, a classroom assessment researcher, to make the case of "validity as an argumentative activity" (p. 68). He goes on to argue that "conceptualizing validity as explicitly a rhetorical activity brings those doing writing assessment and educational measurement to the same table of theory" (p. 68). Inoue turns to sophistic rhetorical theory (the Sophists' articulations of *nomos–physis*) via Plato, Hippias and Antiphon, Thrasymachus, and Protagoras and Prodicus, arguing that sophistic rhetorical theory:

> offers a political sensitivity and philosophy of language that accounts for social contexts and cultural influences on individual readers/judges, allowing validity research to consider individual dispositions to judge in certain ways as consubstantial to larger cultural and historical milieus, creating a complex relationship that can be considered in our validity arguments. (p. 68)

Sophistic rhetorical theory provides Inoue an expansive theoretical framework, and in this way, he is the one writer in this section to dwell deeply in Western rhetorical theory for the theorization of writing assessment. For example, in making the case that "fairness is an investigation of the methods used and the social arrangements and decisions those methods produce (i.e., effects or outcomes)," he draws upon Protagoras: "Protagoras tells us that part of our need for agreement is that each stakeholder has something worthwhile to contribute, some kind of virtue to be tapped. So writing assessment needs more than stakeholder agreement. Writing assessment requires participation" (p. 81).

For Inoue, fairness is not something that is a universal truth; rather, it is "a construction of it, built into it by methods of evidence gathering and judging" (p. 76). Citing Guba and Lincoln (1989), he argues that fairness is "a reflexive method" and "a high level of fairness is achieved when judges/readers "solicit,'

'honor,' and compare various judgments/readings and their 'underlying value structures,' particularly ones that conflict." (p. 76) Again, returning to the sophists, Inoue writes, "For our heuristic, Prodicus calls attention to the healthy conflict within agreement. Agreement is not synonymous with consensus. It is a stance reached through differing readings and judgments, through hard work and *agon*, through disagreement, which could be debate, negotiation, or war" (p. 82). Like fairness, validity, then, "stems from stakeholder ability to participate in and accept decisions from participation" (p. 81).

Inoue's discussion of social implications is most interesting in his analysis of the contributions of classical rhetoric. On one hand, he does not address that the social conditions of classical rhetoric were far from equitable; women, children, foreign residents, and slaves could not participate in Greek democratic activities. On the other hand, readers can see him start to work through ideas about ideology and assessment that he would advance in later publications, such as his work on anti-racist writing assessment (Inoue, 2015). For example, his current work on habits of white language use (2021) is based on the argument that assessment standards are driven by underlying values, values that are based on white supremacy. In his 2007 article, readers can see evidence of his resistance to an ideal model or a standard against all are measured. As Inoue observes:

> Validating writing placement procedures, like validating grades on essays, is also a matter of recognizing clearly how close decisions come to ideal or correct decisions. Validity inquiry that appeals strictly to *physis* typically does not question the dominance of particular values, theoretical frameworks used to make inferences and decisions, or methods for data collection. (p. 73)

Further:

> Viable alternative interpretations and evidence have difficulty competing with dominant frameworks that make up our methods, what constitutes evidence, fairness, and participation in assessments. (p. 71)

And, in conclusion:

> How is the assessment and its results working toward the interests of those being assessed, namely students (and secondarily programs and faculty), and not simply reinforcing the interests of those with power (or those who control the "land" of assessment)? (p. 78)

Because Inoue is most interested in classroom writing assessment, he is attuned to the ways assessment can invite not just participation but also the negotiation of meaning and power. He does not construct the notion of validity, thus, as something about test design. Rather, he writes, validation "might be an inquiry into stakeholder interests and needs, the power created and used, and the assessment's consequences for stakeholder well-being" (p. 78). Such accounting for "individual ways of sensing and judging for those expectations" is an imperative for fairness and agency, rather than domination (p. 88). In the end, Inoue's position rests on agency and that we must have reflexivity in the knowledge to make choices, and the social structure to allow these choices to be made. For Inoue, the really important questions are about cultural hegemony that we reject in public but, in fact, practice within our classrooms. Thus, the really important answers are to be found in the direction of opening doors through our classroom assessment practices.

ETHICAL DIMENSIONS

While O'Neill, Broad, and Boyd, and Inoue were negotiating the relationship of measurement and rhetoric in the early 2000s, by 2016 the field had changed. David Slomp's 2016 article "Ethical Considerations and Writing Assessment" (Chapter 14, this volume) evidences a different interdisciplinary moment. Slomp's article is the introductory article to a special issue of *JWA* on ethics and writing assessment, and the issue contained articles that drew from decolonial theory (Cushman, 2016), civil rights law (Poe & Cogan, 2016), politics (Broad, 2016), and philosophy (Elliot, 2016; Slomp, 2016). In his introduction, Slomp argues that "a theory of ethics compels attention beyond the question of technical competence towards broader questions of social consequences" (p. 97). Slomp's tone suggests that the contributors, of all writing studies researchers, had little interest in commenting on the need for an interdisciplinary landscape of writing studies and measurement. Instead, there was a more direct call to address limitations in measurement:

> Some might question the need for a theory of ethics. After all, the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014) already have defined technical requirements for assessment design and use. Throughout this special issue, however, we argue that technical competence/quality is only one component of ethical practice. Technical quality or feasibility may provide some justifi-

> cation for implementing an assessment practice, but technical feasibility is not equivalent to moral or ethical justification for that practice. (p. 94)

In commenting directly on the field of measurement, Slomp and contributors no longer need to posit a relationship between writing studies and measurement. Instead, there is a move to shape measurement theory itself through humanistic intervention.

Slomp weaves ethics through each of the foundational principles of measurement theory—reliability validity, and fairness. In regard to reliability and the varied forms of evidence accompanying it, he argues that "the demonstration of high degrees of reliability can provide some technical justification for the use of an assessment without addressing deeper ethical questions" (p. 96). In regard to validity, Slomp takes on narrow interpretations of argument-use approaches to validity:

> Validity theorists, themselves, have consistently and explicitly narrowed the breadth of such arguments to focus solely on the uses and interpretations of test results. As such, these arguments are framed as technical ones. . . . We can trust [test scores] because they (a) have been shown to accurately predict future performance; (b) reflect similar scores achieved on similar parallel measures; and (c) accurately reflect the construct the instrument was designed to measure. (p. 96)

The restricted focus of validity arguments, thus, means that questions about construct representation and construct stability raises new questions: "can we defend the use of assessment results for tests that measure constructs we know little about or for where there is little consensus as to what the construct entails?" (p. 97).

In light of the 2014 revised *Standards* which elevated the status of fairness to validity and reliability, Slomp argues that "of the three guiding principles—validity, reliability, and fairness—fairness, with its attention to impacts of assessment practices on individuals, touches most closely on the need for new practices informed by moral philosophy" (p. 100). It is fairness that most attends to social conditions of test use: "In current times, large-scale high-stakes writing assessments may be designed to reflect principles of fairness for individual students while simultaneously being employed to both control and shape education systems" (p. 100).

As Slomp makes clear, none of the three core principles nor the *Standards* is sufficient as an ethical framework for assessment. In response to this gap, Slomp proposes "a theory of ethics for the field of writing assessment, one that advances such a framework toward new conceptualizations that better serve students" (p. 102). He offers six principles based on primary referential frames drawn from diverse stakeholders, exploration of issues related to reliability and validity from

multiple perspectives, adoption of an ecological orientation; emphasis on an integrated approach to evidence, considerations of varied assessment genres, and actionable accountability. From Slomp's perspective, the question is not about the relationship between writing studies and measurement. Here, instruction and assessment as well as evidence of validity, reliability, and fairness are brought together not just through singular referential frames but, rather through ontological, epistemological, and axiological perspectives. The view is interconnected across sites of assessment, across communities of stakeholders, and makes test designers and test-users accountable to "how assessments shape systems of education, and how they impact stakeholders within those systems" (p. 103). In terms of assessment theory, there is no one answer, Slomp suggests. His aim is to trouble those who believe there is, a point he develops in future scholarship (Randall, Poe, & Slomp, 2021; Slomp & Elliot, 2021).

CONCLUSION

According to the CDC's Autism and Developmental Disabilities Monitoring Network "about 1 in 54 children has been identified with autism spectrum disorder (ASD)": ASD occurs in all racial, ethnic, and socioeconomic groups and 4 times more common in boys than in girls (Maenner et al., 2020). Now that neurological diversity is well-known in the field of writing studies (Yergeau, 2018), how will writing assessment change to make it more fair for such students? What lessons from history can we learn about what we can see today and what we cannot see? What can we say about fairness when it is acknowledged that substantial individual differences are part of any assessment? What difference does one person make?

If we think about assessment as situated historically, there are three lessons to be learned from the research contributions in this section of *Considering Fairness and Aspiring to Justice*. First, our understanding of construct—i.e., what is writing?—is always changing (not necessarily evolving); any claims drawn from assessment data are historically contingent. Thus, any claims about fairness must always be tempered by the acknowledgment that our understanding of lack of bias, equity, and justice are always contingent. Second, at each moment in history, assessment technologies and social condition are interlocking—in Brad's case, those technologies were developmental testing, IQ testing, classroom assessment, admissions, and warning systems meant to eliminate failures. The social conditions were social stereotypes, legacies of intergenerational poverty and linguistic discrimination, and whiteness. It was not one test that told a story of Brad's progress, potential, and failure. It was the interlocking of assessments, social conditions, and their consequences. Finally, advances in

assessment technologies are never evenly distributed. We must never assume that any advancement in making assessment fairer will benefit all. There is always an injustice yet to addressed.

In re-reading these contributions spanning over almost two decades, we see the limits of history. Each piece is deeply contextualized within an historical moment, one that provides the exigence of a hopeful future for the authors but also limits what is unseen—advances that stall, historical narratives that are later challenged, and other roads taken. Published just four years apart, O'Neill, Broad, and Boyd, and Inoue were working at a historical moment when the discussion centered around the uses—or not—of measurement theory. In looking to measurement, writing studies researchers selected measurement researchers that seemed to fit the narrative that was needed for writing assessment—a rhetorical approach that invited community engagement.

But in crafting that narrative—an impulse to tell a history of assessment as one of waves (Yancey, 1999)—there was a subsequent erasure of other measurement histories. That history is now part of the racial reckoning that is happening in measurement through projects such as Stafford Hood and Rodney K. Hopson's "Nobody Knows My Name," an endeavor that retrieves "from near obscurity the work of early contributors and pioneering African American scholars who have been excluded from what is taught as the history of educational evaluation research in the United States" (p. 411). In writing about the work of Asa G. Hilliard, for example, Hood and Hopson (2008) write the idea of fairness has been central to such pioneers in the field:

> For nearly three quarters of a century, one issue has guided and driven the work of African American scholars of educational evaluation. Issues of fairness and equity were at the heart of their inquiry in the 1930s when the doctrine of the land mandated so-called separate but equal school systems for children of color. The issues of fairness and equity were central in their investigations of segregated schools during the pre-Brown and supposedly desegregated schools of the post Brown eras. The issue of fairness remains uppermost in our minds today as we investigate our woefully inadequate schools for Black children, other children of color, and children from economically oppressed backgrounds. (p. 413)

By the time that Slomp was writing in 2016, the need to legitimize the field of writing assessment was no longer needed (even if measurement researchers continued to fail in their citations of writing studies scholars; see Behizadeh & Engelhard, 2011). By 2016, researchers like Slomp were less interested in tracing

advances in measurement than in demonstrating how writing assessment research could improve upon the shortcomings of those who continued to believe that standard gauges were the answer to all empirical challenges. By that same time, Inoue, too, had also sharply turned away from measurement as an epistemological orientation to assessment. Today a more tempered view is useful as we watch the standardizers lurch, absorbing notions like culturally and linguistically responsive assessment, but still resisting more radical transformations such as anti-racist assessment, translingual assessment, and neurodiverse validity. In all of it, I wonder what my brother would have felt.

REFERENCES

American Educational Research Association, American Psychological Association, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Behizadeh, N., & Engelhard Jr, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, *16*(3), 189-211.

Boyer, M. (2020). Fairness in educational testing: The role of values in addressing fairness in test purpose, use, and consequences. *Center for Assessment*. https://www.nciea.org/blog/educational-assessment/fairness-educational-testing

Burke, K. (1950). *A rhetoric of motives*. Prentice-Hall.

Cleary, T. A. (1968). Test bias: Prediction of Negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115-124.

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

Dorans, N., & Cook, L. (2016). *Fairness in educational assessment and measurement*. Routledge.

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

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

Educational Testing Service. (2014). ETS standards for quality and fairness. Educational Testing Service. https://www.ets.org/s/about/pdf/standards.pdf

Educational Testing Service. (2016). ETS international principles for fairness review of assessments. Educational Testing Service. https://www.ets.org/s/about/pdf/fairness_review_international.pdf

Gipps, C. & Stobart, G. (2010). Fairness. In B. McGraw, E. Baker, & P. Peterson (Eds.), *International Encyclopedia of Education* (3rd edition, pp. 56-60). Elsevier.

Guba, E., & Lincoln, Y. (1989). *Fourth generation evaluation*. Sage.

Hannant, P., Tavassoli, T., & Cassidy, S . (2016). The role of sensorimotor difficulties in autism spectrum conditions. *Frontiers in Neurology*, *7*, 124. https://doi.org/10.3389/fneur.2016.00124

Hood, S., and Hopson, R. K. (2008). Evaluation roots reconsidered: Asa Hilliard, a fallen hero in the "Nobody Knows My Name" Project, and African educational excellence. *Review of Educational Research*, 78, 410–426. https://doi.org/10.3102/0034654308321211

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning*. Utah State University Press.

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. The WAC Clearinghouse; Parlor Press. https://doi.org/10.37514/PER-B.2015.0698

Inoue A. B. (2021). Above the well: An antiracist literacy argument from a boy of color. The WAC Clearinghouse; Utah State University Press. https://doi.org/10.37514/PER-B.2021.1244

Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement*, *50*, 1-73.

Maenner, M. J., Shaw, K. A., Baio, J., Washington, A., Patrick, M., DiRienzo, M., Christensen, D., Wiggins, L., Pettygrove, S., Andrews, J., Lopez, M., Hudson, A., Baroud, T., Schwenk, Y., White, T., Rosenberg, C., Lee, L., Harrington, R., Huston, M., . . . Dietz, P. M. (2020). Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2016. *MMWR Surveillance Summary*, *69*(4), 1-12. http://dx.doi.org/10.15585/mmwr.ss6904a1external icon

Nuwer, R. (2016). Autism's history holds lessons for today's researchers. *Spectru*m. https://tinyurl.com/mubd2d35

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

Poe, M., Oliveri, M. E., & 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

Randall, J., Poe, M., & Slomp, D. (2021). Ain't oughta be in the dictionary: Getting to justice by dismantling anti-Black literacy assessment practices. *Journal of Adolescent Learning and Literacy, 48*(3), 594-599.

Slomp, D. (2016). An integrated design and appraisal framework for ethical writing assessment. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/4bg9003k

Slomp, D., & Elliot, N. (2021). What's your theory of action? Making good trouble with literacy assessment. *Journal of Adolescent & Adult Literacy*, *64*(4), 468-475.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147-170.

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

Yergeau, M. (2018). *Authoring autism: On rhetoric and neurological queerness*. Duke University Press.

Zieky, M. (2016). Developing fair tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.)., *Handbook of test development* (pp. 81-99). Routledge.