Civil Rights and Writing Assessment: Using the Disparate Impact Approach as a Fairness Methodology to Evaluate Social Impact

by Mya Poe, Northeastern University and **John Aloysius Cogan Jr.**, University of Connecticut School of Law

The Civil Rights Act of 1964 has served as an influential legal framework for addressing intentional (disparate treatment) and unintentional (disparate impact) discrimination. While philosophical and methodological discussions of Title VI and Title VII are well articulated in the legal scholarship, the disparate impact approach--a method for evaluating unintended racialized differences in outcomes resulting from facially neutral policies or practices--remains an underutilized conceptual and methodological framework in assessment literature. In this article, we argue that the burden-shifting heuristic used by entities such as the Office for Civil Rights to redress disparate impact is a valuable approach in evaluating fairness of writing assessment practices. In demonstrating an application of the burden-shifting approach at one university writing program, we discuss the value of the proposed integrative framework and point to remaining questions regarding sampling concerns--group identification, group stability, and intersectionality.

On June 11, 1963, U.S President John F. Kennedy delivered what has become known as the Civil Rights Address, a speech given the evening after Alabama National Guardsmen were sent to the University of Alabama to "carry out the final and unequivocal order of the United States District Court of the Northern District of Alabama" that required the university to admit "two clearly qualified young Alabama residents who happened to have been born Negro." In Kennedy's address regarding the admission of Vivian Malone and James Hood to the University of Alabama, he invoked the ideals of human rights, tolerance, reciprocity, and color-blindness. He called the issue of equal rights a "moral issue," an issue that every American should embrace because of its connections to the founding principles of American democracy:

> I hope that every American, regardless of where he lives, will stop and examine his conscience about this and other related incidents. This Nation was founded by men of many nations and backgrounds. It was founded on the principle that all men are created equal, and that the rights of every man are diminished when the rights of one man are threatened.

Kennedy also invoked the notion of standards in his use of the phrase "clearly qualified." In doing so, he signaled that Malone and Hood were not being given special privileges because they were African American. By the university's admissions standards, they were qualified—"clearly qualified"—for admission.

Kennedy went on in his speech to trace the relationship between opportunity, talent, and motivation:

> As I've said before, not every child has an equal talent or an equal ability or equal motivation, but they should have the equal right to develop their talent and their ability and their motivation, to make something of themselves. (June 11, 1963)

For Kennedy, access—the right to develop one's talent—was more important than the actual talent one possessed. Measurement of ability was secondary to equitability.

Kennedy's vision would become codified after his death in the Civil Rights Act of 1964. The Act would advance not just a moral dictum for eliminating discrimination but also a legal framework for actionable standards—a framework that outlawed barriers to access through intentional as well as unintentional discrimination. Specifically, in identifying unintentional discrimination, what would become known as "disparate impact"—"facially neutral policies that are not intended to discriminate based on race, color, or national origin, but do have an unjustified, adverse disparate impact on students based on race, color, or national origin" (Department of Education, 2014, p. 8)—the Civil Rights Act of 1964 has given us a framework for evaluating and remedying barriers to access that are not immediately visible.

Today, in higher education the barriers set through placement and proficiency testing can be enormous. The number of students whose lives are affected by our decisions to deny them access to first-year courses is startling. For example, in 2006 in the California higher education system, 30% of students in the university system, 60% in the state system, and 90% in the community college system required remediation (Murray, 2008). Nationally, approximately 20% of students entering four-year colleges and 50% of students entering community college require remediation (Complete College America, 2012). And the numbers for students of color are even more sobering. African American students are placed in remedial classes at rates of almost 40% for four-year colleges and 67% for two-year colleges. Hispanic students are placed at rates of 21% and 58% respectively while white students are placed at rates of 14% and 47% (Complete College America, 2012, p. 6).[1]

When it comes to course completion, again, the numbers for students of color are dismal. Almost 70% of African American students in four-year colleges and more than 85% of African American students in two-year colleges did not complete remedial and associated college-level courses within two years. Hispanic and white students faired only a bit better at approximately 64% and 76%, respectively (Complete College America, 2012, p. 8). And graduation rates? They are adversely affected as well. While

nationally, the overall six-year graduation rate for students enrolled in four-year colleges is well over one-half (55.7%), the graduation rate falls by over one-third to 35.1% for students required to complete remedial and additional coursework. The same effect can be seen in the graduation rate at two-year colleges. The overall three-year graduation rate at those schools is 13.9%, but drops by nearly one-third to 9.5% for students required to complete remedial and additional coursework (Complete College, 2011, p. 14).

In identifying students who need additional help for writing, courses like basic writing have an important place in higher education. Approaches ranging from studio models (Grego & Thompson, 1995, 2007) to stretch programs (Glau, 1996) to accelerated instruction (Adams et al., 2009) have all been innovations to better support students enrolled in basic writing. Without such courses, many students would find themselves without the support they need to develop college-level writing practices. More importantly, corequisite classes like studio, stretch, and accelerated basic writing have been shown to work; students who enrolled in single-semester, corequisite English courses typically succeeded at "twice the rate of students [enrolled] in traditional prerequisite English courses" (Complete College America, 2015a, n.p.) Yet, corequisite options remain the exception at many institutions where basic writing typically does not carry college credit toward graduation and students must pass an exit exam to matriculate into first-year writing (Isaacs, forthcoming, p. 129).

Ultimately, students of color and multilingual students are the most likely to face the negative consequences of remediation (Sternglass, 1997; Soliday, 2002). Institutional writing assessment practices are often selected without regard to their effects on diverse student populations (Lioi & Merola, 2012; Elliot et al., 2012), human readers and machines alike can respond quite differently to identity markers in essays (Lindsay & Crusan, 2011; Marefat & Heydari, 2016; Shermis, Lottridge, & Mayfield, 2015), and scoring procedures can yield quite different predictive results (Wilson et al., 2016). If test design and curriculum are so fraught with questions about equitability, are equitable outcomes simply comparable test scores, as has been the assumption behind legislation the recent reauthorization of the Elementary and Secondary Education Act of 1965 by the Every Student Succeeds Act (2015)? What if test scores reflect unequal opportunity to learn—i.e., the conditions that promote learning for students? And, finally, what is the relationship between fairness and equity?

In making this argument, we are extending the work previously published with our colleagues (Poe, Elliot, Cogan, & Nurudeen, 2014) in which we demonstrated the use of the Department of Education Office for Civil Rights (OCR) methodology to demonstrate its viability for writing program self-study. Here, we deepen our previous work to discuss the conceptual value of disparate impact as part of an ethical framework for writing assessment. We begin with a discussion of fairness as currently found in the *Standards for Educational and Psychological Testing* (*Standards*) (AERA, APA, & NCME, 2014) and in the measurement literature. We then discuss the various means by which discrimination has been addressed through the courts to frame our discussion of disparate impact. After a detailed discussion of the Civil Rights Act and Title VI, we then explain the disparate impact approach as applied through the OCR. Applying the OCR "burden-shifting approach" in a writing assessment case, we discuss the methodological questions that remain unaddressed through the disparate impact approach as well as identify its conceptual and methodological potential.

Two caveats here are important before proceeding: First, we are not advancing a legal argument for or against the use of disparate impact theory (Braceras, 2005). We are simply arguing that the disparate impact approach, which has been refined and has withstood numerous challenges for more than 50 years to determine when societal action was needed to reassess the interpretation of outputs and remedy the unequal distribution of inputs in a variety of institutional settings, can be a valuable tool to assess the differential effects of assessment practices. Furthermore, the Department of Education Office for Civil Rights' "burden-shifting approach" is a valuable heuristic—akin to a validation study—for remedying differential effects. Second, for the sake of simplicity, our discussion in this article is limited to claims of racialized differences. The disparate impact approach, however is flexible and has the capacity to identify disparate impact across other group identities (e.g., sexual orientation) (Department of Justice, 2015).

1.0 Shifting Conceptions of Fairness

In lieu of a sustained coherent discussion about the history of fairness frameworks in measurement, the authors of the *Standards* attempted to provide a technical framework for fairness by linking it to validity:

> The validity of test score interpretations of intended use(s) for individuals from all relevant subgroups. A test that is fair minimizes the construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some individuals. (AERA, APA, & NCME, 2014, p. 219)

What may be concluded from the current edition of the *Standards* is that the current view of fairness rests on access to constructs measured (e.g., how individuals respond in testing contexts and thus offering appropriate modifications or adjustments) and score interpretation (e.g., disaggregating scores to determine group differences). It is a view of fairness located in a moment in time—at a point of access or *in témpore* score interpretation, not as an ongoing decision-making process, which would be consistent with current views of validity. This view of fairness, also, does not locate it within a theory of action, such as found in through-course validity arguments (Bennet, Kane, & Bridgeman, 2011). In the end, the authors of the *Standards* left the larger challenge for fairness

—the relationship of "opportunity to learn" to social consequences—relatively untouched. Such omission is not unexpected given the measurement community's conventional views on opportunity to learn. As Heartl, Moss, Pullin, & Gee (2008) argued, prevailing psychometric conceptions of opportunity to learn locate knowledge "inside the heads of individual learners, privileging symbolic representation over embodied experience, and relegating the social dimensions of learning…to the role of background or context in the business of measuring learning outcomes" (p. 3).

This is not to say, however, that the measurement community has always been limited by an epistemological separation of innate ability and social context or has not seriously engaged with issues related to ethics and fairness. For example, in the late 1960s the American Psychological Association established the Task Force on Employment Testing of Minority Groups. The task force was comprised of measurement researchers like Samuel Messick, who would go on to champion consequential validity, and led by Brent Baxter, an industrial psychologist who worked for Prudential Insurance Company and would later become Vice President of the American Institutes for Research. The committee published its findings in "Job testing and the disadvantaged" (APA, 1969)—a report framed in a way that is consistent with Kennedy's vision of equality of opportunity:

> In an ideal world …Each person would use his capabilities in the most productive and self-enhancing fashion, and his society thereby would make the wisest and most humane use of its manpower resources. Such a goal is not easily realized. Its attainment may be blocked sometimes by the personal maladaptive tendencies of the individual. More generally, however, it is society that often thwarts the matching between an individual's capabilities and his vocational role. (APA, 1966, p. 637)

The report examined "the chain of events that can lead to the inappropriate use of manpower and unfair and self defeating personnel practices" (APA, 1966, p. 637). While the authors argued that knowledge-based tests are "free of bias," they also argued that aptitude testing is "a more subtle and complex issue" (APA, 1966, p. 640) because of "cultural deprivation" (an unfortunate choice of wording), "test-induced anxiety," "unfairness of test content," "improper interpretation of test scores," and "lack of content relevance" (APA, 1966, pp. 640-642). Thus, in outlining the various dimensions by which aptitude tests may misrepresent an examinee's actual abilities, the authors of the Baxter report pointed to the flawed logic of standardization—that consistency is equivalent to fairness.

In 1976 a special issue of the *Journal of Educational Measurement* (*JEM*) was devoted to the topic of bias. As Jaeger (1976) wrote in the introduction, "Attempts to advise the U.S. Department of Justice on an appropriate definition of 'fair' selection have resulted in 'an agreement to disagree'" (p. 1), resulting in a tenuous statement in the 1974 edition of the *Standards* regarding the definition of fairness: "It is important to recognize that there are different definitions of fairness, and whether a given procedure is or is not fair may depend upon the definition accepted" (AERA, APA, & NCME, 1974, p. 44). The goal of 1976 *JEM* issue, then, seemed to provide some guidance to subsequent editions of the *Standards* and educate the practitioner community that fairness was no longer simply "selection . . . based on the predicted criterion" (Sawyer, Cole, & Cole, 1976, p. 59). This goal was achieved under the guise of giving authors who had contributed to earlier fairness models "an opportunity to bring their ideas up to date, and to comment on the [new fairness] model proposed [in the lead article to the special issue] by Petersen and Novick" (Jaeger, 1976, p. 1).

In their article, Petersen and Novick attempted to correct for faulty judgments in "culture-free selection" and group parity models, such as the regression model that equates optimal prediction for lack of bias, that end up sanctioning "the very discrimination they seek to rectify" (1976, p. 5, p. 28). The article received mixed reviews. While Cronbach in the same *JEM* special issue praised Petersen and Novick, he also noted, "most of the attention has been given to the simplest of payoff matrices, uniform for all groups, and to single-stage selection. In time, it will be necessary to derive indices of fairness that reflect more complex matrices" (1976, p. 40). In another article in the *JEM* special issue, Linn advocated for a "decision-theoretic" approach. The decision-theoretic approach, he argued, allowed for public scrutiny and debate about value judgments. Linn went on to argue that such an approach, one that is "a way of formalizing the judgments and observing the consequences," "makes the process of attaching values to different outcomes a political one [rather than purely a technical one], which is what it should be" (1976, p. 56). In the end, the authors of the 1976 *JEM* special issue seemed resigned, as Breland and Ironon concluded, that "the solution to the broad social dilemma [of inequality] is not to be found in psychometric models" (1976, p. 98).

By the 1980s, Cronbach and Messick were both arguing that social consequences were related to validity. For example, in his 1989 article, "Meaning and Values in Test Validation: The Science and Ethics of Assessment," Messick wrote that social consequence was integral to a unified theory of validity: "The key issues of validity are the meaning, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of the social consequences of their use" (p. 5). Yet, Cronbach and Messick disagreed as to the reach of social consequences. While Cronbach (1988) argued that "tests that impinge on the rights and life chances of individuals are inherently disputable" (p. 6), Messick argued:

> If the adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardized. If the social consequences cannot be so traced—or if the validation process can discount sources of test invalidity as the likely determinants, or at least render them less plausible—then the validity of the test use is not

> overturned (1989, pp. 88-89)

In short, Messick was worried about consequences that strayed too far from a test's construct meaning.

Through the 1990s and into the 2000s, various articles appeared that wrestled with the degree of social consequences in relation to validity and fairness (e.g., Cole & Zieky, 2001; Gallagher, Bridgeman, & Cahalan, 2002; Kane, 2012; Langenfeld, 2005; Lu & Suen, 1995). In addition to the collection *Fairness and validation in language assessment* from the 19th Language Testing Research Colloquium (Kunnan, 2000), one of the notable publications on the subject was the Moss et al. 2008 collection *Assessment, equity, and opportunity to learn*, both of which squarely took on the issue of social justice. For example, in her contribution to the Moss et al. collection, Pullin (2008) wrote:

> Equally significant [to measuring outcomes] are the implications of assessment for equity and social justice, insuring that *all* students, particularly those most at risk of educational failure, are the beneficiaries of an effective *opportunity to learn* (OTL) meaningful content . . .This leads to a dramatically new perspective on OTL, not in terms of content covered and scores attained, but instead based on a more complex view centered on aspects of learning activities and the role of assessment as part of the learning environment. (p. 334).

Recent research that wrestles with the question of whether fairness should be subsumed under validity or strive for broader social justice goals includes Xi (2010) on comparable validity, Mislevy et al. (2013) on universal design, Solano-Flores (2002) on cultural validity, and Steele & Aronson (1995) on stereotype threat. Following Kane (2006), Xi has advanced an argument of fairness as "as comparable validity for all *relevant* groups" (p. 147). Working in the field of language testing, Xi's approach includes adding a corresponding fairness claim to each validity claim: "the fairness argument consists of a series of rebuttals that may challenge the comparability of scores, score interpretations, score-based decisions and consequences for sub-groups" (Xi, 2010, p. 157). Such an approach has also been used by Slomp, Corrigan, & Sugimoto (2014) to evaluate consequences.

One concern about the marriage of fairness and validity is whether an argument-based approach to fairness via validity is too unwieldy. As Borsboom (2005) has pointed out, the expansion of validity theory in-and-of itself has resulted in an unwieldiness in practice (a point demonstrated by Chapelle, Enright, & Jamieson, 2010):

> In the past century, the question of validity has evolved from the question whether one measures what one intends to measure from the question whether the empirical relations between test scores match theoretical relation in a nomological network (Cronbach and Meehl, 1955), to the question whether interpretations and actions based on test scores are justified—not only in the light of scientific evidence, but with respect to social and ethical consequences of test use (Messick, 1989). Thus, validity theory has gradually come to treat every important test-related issue as relevant to the validity concept, and aims to integrate all these issues under a single header. In doing so, however, the theory fails to serve either the theoretically oriented psychologist or the practically inclined tester. (Borsboom, 2005, pp. 149-150).

We take Borsboom's point to heart. If fairness in writing assessment design is to be achievable, it must appeal to both the theoretically-oriented writing researcher and the writing program administrator who needs to easily gather and present data to a wide range of stakeholders, often under very limited time constraints.

Another recent approach to fairness has been through universal design (i.e., access). Universal design is based on the premise that careful definitions of the construct to be measured can minimize test taker characteristics that interfere with score interpretation, or as Mislevy et al. (2013) explained, "deliberately varying aspects of an assessment for students to enable each student to access, interact with, and provide responses to tasks in ways that present minimal difficulty" (p. 122). Universal design is important because it challenges existing approaches that attempt to "retrofit" assessment to diverse student population—i.e., design a "color-blind" test and then account for diverse response processes (Mislevy et al., 2013, p. 137). Yet, while universal design acknowledges differences among test takers that may result in the misinterpretation of scores, thus aligning it more closely with socio-cultural perspectives (Behizadeh, 2014), it remains focused on access to construct representation for the purposes of score interpretation. Moreover, it assumes that we can know enough about latent responses to validate claims (i.e., latent variables are identifiable), that latent variables are stable within groups and for individuals (i.e., individual learning and development is ignored)[4], and that there is homogeneity within groups (i.e., that racial/ethnic groups are sufficiently homogenous in cognitive and social profiles). In the end, while latent variable analysis may be useful for identification of genre features (e.g., what are common features of proposals), it can be very easily abused in essentializing writing performances of identity groups.

Cultural validity, likewise, is interested in "the socio-cultural influences that shape student thinking and the ways in which students make sense of …items and respond to them" (Solanes-Flores & Nelson-Barber, 2001, p. 555). [5] While the roots of universal design research stem from test accommodations for disabled students (Americans with Disability Act), cultural validity research stems from studies of linguistically diverse students. Like universal design, it neither accounts for historical conditions nor the unintended discrimination that arises from those conditions. It also inadvertently ties linguistic identity to racial/ethnic identity,

assuming that latent variables are universal (or universal enough) across a group as to be meaningful for the purposes of designing fair assessment practices.

Stereotype threat theory, which is not the same as test anxiety, was developed to account for the lasting effects of discrimination. Stereotype threat postulates that students who identify with a particular domain (e.g., math) falter in performance when they struggle to overcome misconceptions about their abilities in that domain (e.g., women are bad at math). Stereotype threat research has been extended to a number of conditions (e.g., race, socioeconomic status, gender) (Nguyen & Ryan, 2008) and has been usefully applied in classroom conditions (Aronson, Fried, & Good, 2002; Cohen, Steele, & Ross, 1999). Its applicability to test design has been limited because it is not clear what methodologies are to be developed from it for the purposes of assessment (Good, Aronson, & Inzlicht, 2003; Striker, 2008; Striker & Ward, 2004; Walker & Bridgeman, 2008; Yaeger & Walton, 2011). Nonetheless, its implications for assessment are on the horizon. For example, research by Walton and Spencer (2009) has pointed out that the ability of stereotyped students is latent, thus "underestimated by their level of prior performance" (p. 1133) and that "threat" may actually increase "at each rung of the educational ladder" (p. 1133). In a series of studies, they found that underestimation of intellectual ability was the result of psychological threat, but that "psychological treatments can recover much of this otherwise lost human potential" (Walton & Spencer, 2009, p. 1137). Walton and Spencer (2009) argued, "To close achievement gaps, it is necessary both to eradicate psychological threats embedded in academic environments and to remove other barriers to achievement including objective biases, the effects of poverty, and so forth" (p. 1137).

In the end, although the current issue of the *Standards* suggests otherwise, the assessment community has long wrestled with questions of fairness in testing. In what follows, we seek to add to that conversation by drawing on the disparate impact analysis framework. Before continuing with our discussion, we explain the legal context from which the method was derived and how the method has been used. In the following section, we begin by setting forth the various legal standards—constitutional, statutory, and regulatory—through which racial discrimination has been addressed. This contextualization is critical in understanding the impediments faced by claimants alleging unintentional discrimination and theorization difficulties faced by courts addressing such claims. This background also situates the disparate impact approach and its burden-shifting methodology among the field of legal approaches to racial discrimination. We then discuss the history of the Civil Rights Acts, including Title VI: Nondiscrimination in Federally Assisted Programs before concluding with a discussion of the OCR process—the process used at all federal agencies—to address complaints.

2.0 Legal Pursuit of Discrimination Claims

While federal laws prohibiting racial discrimination date back to the post Civil War era, the century that followed the Civil War saw only limited progress in ending racial discrimination. In attempting to address continued and pervasive racial discrimination, Presidents Kennedy and Johnson sought to lay out legal frameworks that complemented constitutional rights and augmented gaps in existing state and federal statutes and regulations.[6] For example, in addition to the Civil Rights Act, the Voting Rights Act of 1965, was enacted to prevent and remedy racial discrimination in voting, and the Fair Housing Act (1968), was enacted to prohibit discrimination in real estate sales, rental, lending, insurance, and other related services based on race, color, sex, religion, and national origin (with familial status and handicap added later).

Specifically, in an educational context, discrimination can be challenged through various avenues, including constitutional, statutory, and regulatory paths: (1) under the Equal Protection Clause of the Fourteenth Amendment to the United States Constitution, (2) under Title VI of the Civil Rights Act of 1964 (or with complaint to the U.S. Department of Education based on Title VI regulations), (3) under 42 U.S.C. § 1983, (4) under state constitutional provisions, and (5) under state constitutional and statutory/regulatory anti-discrimination laws. Table 1 summarizes these avenues with a state example taken from a single state, New Jersey.[7], [8]

Table 1: Comparison of Federal and State of Laws Against Discrimination

	Federal			State (NJ)	
	Constitutional	**Statutory/ Regulatory Claims**		**Constitutional**	**Statutory/ Regulatory Claims**
	Equal Protection Clause of the 14th Amend.	Title VI of the Civil Rights Act of 1964[a]	42 U.S.C. § 1983	Equal Protection under NJ state constitution	NJ Law Against Discrimination
Institution Type					
Applicable Against Public Institutions	Yes	Yes	No	Yes	Yes
Applicable Against Private Institutions	No	Yes	No	No	Yes[b]
Private Claim Available					
Intentional Discrimination	Yes	Yes	No	Yes	Yes
Disparate Impact	No	Yes, but may only be enforced by OCR.	Not likely	No	Yes

[a] Applies only to recipients of federal funds.
[b] The N.J. Law Against Discrimination does not apply to private religious educational institutions.

Each approach has non-obvious limitations with respect to disparate impact claims. For example, a practitioner might assume that the most obvious legal avenue for a discrimination claim would be the Equal Protection Clause of the Fourteenth Amendment to the U.S. Constitution. The Equal Protection Clause states, "no State shall . . . deny any person within its jurisdiction the equal protection of the laws" (U.S. Const. amend. XIV, § 1). But the Equal Protection clause is subject to two significant limitations. First, it is only applicable to state, not private, action. Thus, while a public university's policy that expressly discriminates based on race would fall within the ambit of the Equal Protection Clause, the same blatantly discriminatory behavior undertaken by a private university would not involve state action and therefore would not violate the Equal Protection Clause (*Powe v. Miles*, 1968). Moreover, the fact that a private school receives government funding and is heavily regulated by public authorities does not render the school a state actor for the purposes of the Equal Protection clause (*Rendell-Baker v. Kohn*, 1982). Second, the Equal Protection clause does not apply to disparate impact claims. The Supreme Court has made clear that the Equal Protection Clause only prohibits actions that can be shown to constitute intentional discrimination (*Washington v. Davis*, 1976).

Some commentators have suggested that it might be possible to bring a private discrimination lawsuit based on one federal statute (Section 1983 of Title 42 of the U.S. Code) to make a disparate impact claim under another federal statute (Section 602 claim under Title VI of the Civil Rights Act of 1964) (Kidder & Rosner, 2002). Section 1983 does not create rights. Instead, as part of the Civil Rights Act of 1871, it was designed as a vehicle to redress violations of federal Constitutional and statutory rights to combat Reconstruction Era racial violence by the Ku Klux Klan and other White supremacists in the Southern states. In theory, a plaintiff could sue under Section 1983 to redress a violation of his or her federal civil rights by a government official. However, the Supreme Court has never squarely addressed this issue, although federal circuit courts have. Those decisions are split as to whether Section 1983 may be used for disparate impact claims. For example, the Third Circuit (covering Delaware, New Jersey, and Pennsylvania) has ruled that Section 1983 may not be used to enforce disparate impact regulations promulgated under Title VI (*South Camden Citizens in Action v. New Jersey Department of Environmental Protection*, 2001). Likewise, the Sixth Circuit (covering Tennessee, Ohio, Michigan, and Kentucky) and the Ninth Circuit (covering California, Oregon, Washington, Nevada, Montana, Idaho, Arizona, Alaska, and Hawaii) have also ruled that Section 1983 may not be used to enforce disparate impact regulations promulgated under Title VI (*Wilson v. Collins*, 2008; *Save Our Valley v. Sound Transit,* 2003). However, the Tenth Circuit (covering Colorado, Kansas, New Mexico, Oklahoma, Utah, and Wyoming), has indicated that Section 1983 may be used to enforce disparate impact regulations promulgated under Title VI (*Robinson v. Kansas,* 2002). Yet, even in those areas where a circuit court has not explicitly ruled out the use of Section 1983 to enforce disparate impact regulations, the likelihood of a court allowing such a claim is slim (Daly, 2006; Black, 2002). The bottom line regarding the use of Section 1983 to enforce a disparate impact claim under disparate impact regulations is that the standard is applied inconsistently by intermediate-level appellate courts and may not withstand a Supreme Court challenge, thus leaving no national standard.

Finally, in addition to federal laws, some states provide a remedy for disparate impact discrimination (e.g., Or. Rev. Stat. § 659.850 (West Supp. 2015); 740 Ill. Comp. Stat. § 23/5 (2004); Cal. Gov't Code § 11135 (West Supp. 2015); Cal. Gov't Code § 11139 (West Supp. 2015)). However, state constitutions and laws vary as to whether disparate impact is available and if so, how it is applied.

As explained below, Title VI of the Civil Rights Act of 1964 remains the primary legal avenue for addressing claims of disparate

impact for federally-funded programs and facilities and the OCR burden-shifting approach remains the most viable conceptual and methodological guidance from which an approach to fairness in assessment may be developed.[9]

2.1 The Civil Rights Act of 1964 and Disparate Impact

Signed by President Lyndon B. Johnson, the Civil Rights Act of 1964 was landmark legislation prohibiting discrimination in housing, employment, and education. The preamble to the Act states that its purpose is:

> To enforce the constitutional right to vote, to confer jurisdiction upon the district courts of the United States to provide injunctive relief against discrimination in public accommodations, to authorize the Attorney General to institute suits to protect constitutional rights in public facilities and public education, to extend the Commission on Civil Rights, to prevent discrimination in federally assisted programs, to establish a Commission on Equal Employment Opportunity, and for other purposes.

The Act extends the protections granted in the Fourteenth Amendment of the Constitution. Through eleven titles or sections, the Act addresses discrimination in the use of public facilities and accommodations, access to educational facilities, employment hiring and promotion, and voting rights. The Act also establishes various mechanisms for addressing social inequality, including paying for training institutes for teachers, conducting empirical studies to assess ongoing discrimination in educational settings and voter registration, and permitting the Attorney General to initiate legal proceedings in discrimination cases. Finally, the Act sets rules for hearings conducted by the Commission on Civil Rights, which had been established under the Civil Rights Act of 1957, and establishes the Community Relations Service through the Department of Commerce.

The Act addresses discrimination along multiple axes: location, funding, and types of discrimination. On one axis, the Act targets locations of discrimination, ranging from such social institution as schools and hotels. For example, Title II: Injunctive Relief Against Discrimination in Places of Public Accommodation states individuals should have “full and equal enjoyment of the goods, services, facilities, and privileges, advantages, and accommodations of any place of public accommodation. . . without discrimination or segregation on the ground of race, color, religion, or national origin” (1964, §201).

On another axis, the Act targets funding mechanisms, specifically recipients that receive federal funds. Title VI: Nondiscrimination in Federally Assisted Programs, §601, for example, provides:[10]

> No person in the United States shall, in the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving Federal financial assistance. (42 U.S.C. § 2000d)

From a theoretical point of view, what is striking about the Act is the way it captures discrimination (See Perry, 1991 for a useful review of discriminatory purpose theories). The Act acknowledges that both intent (“disparate treatment”) and lack of attention (“disparate impact”) can result in discrimination. This conceptual framework has been instrumental in the shaping the uptake of the Act in Supreme Court decisions. For example, as Chief Justice Berger wrote in the decision for *Griggs v. Duke Power Company* (1971) case:

> [Although] the Company had adopted the diploma and test requirements without any “intention to discriminate against Negro employees.” (420 F.2d at 1232). . . . good intent or absence of discriminatory intent does not redeem employment procedures or testing mechanisms that operate as “built-in headwinds” for minority groups and are unrelated to measuring job capability.

In this way, the Act’s architects saw discrimination as located not only in individual action but also in institutional and social practices. Past discrimination was linked to current effects (“built-in headwinds”), thus acknowledging the temporal aspects of discrimination. In other words, the effects of racist policies and actions—including assessment policies and practices—may not be known until after their effects have occurred.

2.1.1. Title VI: Nondiscrimination in federally assisted programs. Title VI of the Civil Rights Act of 1964 provides: “No person in the United States shall, on the ground of race, color, or national origin . . . be denied the benefits of, or be subjected to discrimination under any program . . . receiving Federal financial assistance” (42 U.S.C. § 2000d). Title VI regulations thus prohibit recipients of federal funds from engaging in practices that “utilize criteria or methods of administration which have the effect of subjecting individuals to discrimination because of their race, color, or national origin” (34 C.F.R. § 100.3(b)(2)). The statute allows for the possibility that federal funds can be denied to a federal grantee—private and public universities—that discriminates (42 U.S.C. § 2000d-1(1)).

Title VI was the most controversial provision in the Act because its vast regulation of the use of public funds. In calling for the

enactment of Title VI, Kennedy stated:

> Simple justice requires that public funds, to which all taxpayers of all races contribute, not be spent in any fashion which encourages, entrenches, subsidizes, or results in racial discrimination. *Direct discrimination by Federal, State, or local governments is prohibited by the Constitution. But indirect discrimination, through the use of Federal funds, is just as invidious; and it should not be necessary to resort to the courts to prevent each individual violation* [emphasis added]. (H.R., 1963)

By targeting discrimination through the federal government's spending powers (Watson, 1990), Kennedy was prescient in understanding that it was insufficient to address discrimination only in existing social institutions.

Unlike the Equal Protection clause of the U.S. Constitution, which only prohibits intentional discrimination by a state actor, Title VI of the Civil Rights Act of 1964 applies to intentional and non-intentional discrimination by state and private actors. However, Title VI does not define what constitutes "discrimination" and does not specify whether the statute includes only intentional discrimination or whether it also reaches more subtle forms of discrimination, such as those that produce racialized disparate effects (Abernathy, 1981; Watson, 1990). Although Congress debated the issue of whether Title VI banned only segregation or extended to de jure discrimination, it never resolved the question. In 2001, however, the Supreme Court, in its *Alexander v. Sandoval* decision, provided some guidance, and in doing so severely restricted disparate impact claims under Title VI.

While the Supreme Court upheld disparate impact in the *Sandoval* case, it foreclosed the ability of private litigants to initiate Title VI disparate impact suits in federal court as it determined Title VI does not create a private right of action (that is, an ability for private, non-governmental actors to initiation legal action) for disparate impact claims. The Court did, however, leave open the possibility of enforcement through agency proceedings (Abernathy, 2006). This means that private parties may file disparate impact complaints with federal agencies, such as the Department of Education, which have the power to investigate, review, and revoke federal funds pursuant to Title VI (42 U.S.C. § 2000d-2). Thus, while the *Sandoval* decision precluded a private lawsuit to enforce a disparate impact claim under Title VI, someone aggrieved by the discriminatory impact of a test can still file a complaint with the U.S. Department of Education alleging disparate impact.

2.1.2 Investigation of disparate impact claims via U.S. Department of Education Office of Civil Rights. In a 2014 letter, the Department of Education noted,

> School districts that receive Federal funds must not intentionally discriminate on the basis of race, color, or national origin, and must not implement facially neutral policies that have the unjustified effect of discriminating against students on the basis of race, color, or national origin. (Department of Education, p. 5)

Individuals can file complaints with the OCR alleging that an institution's assessment practices have a Title VI discriminatory effect on the basis of race (Department of Justice, 2001; Department of Education, 2012).

When investigating complaints of disparate impact, the OCR will undertake a three-step inquiry as outlined in Table 2.

Table 2: OCR's Process for Complaint Inquiry

Step	Question
1.	Does the school district have a facially neutral policy or practice that produces an adverse impact on students of a particular race, color, or national origin when compared to other students?
2.	Can the school district demonstrate that the policy or practice is necessary to meet an important educational goal? If the policy or practice is necessary to serve an important educational goal, then OCR would continue to Step 3.
3.	Are there comparably effective alternative policies or practices that would meet the school district's stated educational goal with less of a discriminatory effect on the disproportionately affected racial group; or, is the identified justification a pretext for discrimination? (Department of Education, 2014, p. 8)

2.1.2.1 ***Step 1—Does the school district have a facially neutral policy or practice that produces an adverse impact on students of a particular race, color, or national origin when compared to other students?*** The first requirement for making a Title VI disparate impact claim is evidence of a discriminatory effect on minority applicants. As the Department of Education letter (2014) makes clear, "Applying this disparate impact framework, OCR would not find unlawful discrimination based solely upon the

existence of a quantitative or qualitative racial disparity resulting from a facially neutral policy" (p. 8). The effect or impact of such policies must be demonstrated through a multi-phase inquiry.

Courts have traditionally relied on a four-step process method to assess impact:

> (a) calculate the pass rate for each group, (b) observe which group has the highest pass rate, (c) calculate measures of impact by comparing the pass rate for each group with that of the highest group, (d) and observe whether the difference in pass rates is substantial. (Fassold, 2000, p. 460-461)

In other words, test score difference alone does not constitute a case of disparate impact. There must be evidence of impact, as well. The courts have not relied on a single measure to assess "impact," but four common methods include the Hazelwood rule, Shoben rule, a rule of practical significance, and the four-fifths rule.

As Fassold (2000) explained, "The Hazelwood rule is based on the binomial distribution taking into account the standard deviation of a binomial event" (p. 42). Used in cases such as *Castañeda v. Partida* (1977), the Hazelwood rule is appropriate where (1) there are only two possible outcomes—e.g., the selection of an African American candidate or a white candidate from a pool of applicants—and (2) where the observed number is greater than two to three times the standard deviation of the expected value.

The Shoben rule is similar to the Hazelwood rule in that it relies on statistical significance. Under the Shoben rule, independence is assumed in that the performance of one individual is not dependent on the performance of another individual. The rule also assumes that sample size is sufficiently large and representative of the population. If these three conditions are met with a 95% confidence interval, "A difference or '*Z*' value greater than 1.96 standard deviations is ordinarily sufficient to support a finding of adverse [racial] impact" (*Richardson v. Lamar County Board of Education*, 1989, p. 816).

The four-fifths rule and the rule of practical significance are complementary approaches. Under the four-fifths rule, disparate impact is found when the effects of a policy or practice have a pass rate of less than 80%, or four-fifths, on a particular race versus the rate of effects on the reference group (West-Faulcon, 2009). Because the four-fifths rule does not take sample size into consideration, it is sometimes complemented with the rule of practical significance. The rule of practical significance is a measure of magnitude of difference where statistical significance can be determined because of sample size (Fassold, 2000, p. 464).

As obvious from the discussion above, impact is a statistical argument–observed value two to three times the standard deviation of the expected value, a *Z* value greater than 1.96 standard deviations, or pass rates of less than 80%. More importantly, while the statistical determination of disparate impact is valuable, statistical analysis alone does not probe the underlying arguments for differential outcomes. It also does not suggest what remedies should be put into place to address adverse impact or how that process might unfold.

2.1.2.2 ***Step 2—Can the school district demonstrate that the policy or practice is necessary to meet an important educational goal?*** In conducting the second step of this inquiry, the university is given the opportunity to rebut the evidence of discriminatory effect by demonstrating that the criterion that resulted in the impact is required by educational necessity. OCR would consider both the importance of the educational goal and the tightness of the fit between the goal and the policy or practice employed to achieve it. If the policy or practice is not necessary to serve an important educational goal, OCR would find that the school district has engaged in discrimination. If the policy or practice is necessary to serve an important educational goal, then OCR would continue to Step 3.

2.1.2.3 ***Step 3—Are there comparably effective alternative policies or practices that would meet the school district's stated educational goal with less of a discriminatory effect on the disproportionately affected racial group; or, is the identified justification a pretext for discrimination?*** If the answer to either question is "yes," then OCR would find that the school district had engaged in discrimination. In other words, if the defendant university successfully demonstrates that the racialized disparate impact of its policy is educationally justified, the institution is still liable for violating Title VI if there is evidence that a less discriminatory alternative exists to the challenged criterion. If no, then OCR would likely not find sufficient evidence to determine that the school district had engaged in discrimination (Department of Education, 2014, p. 8).

Upon conclusion of the process, OCR process begins with efforts at voluntary compliance first. When such cases fail, the OCR can initiate an enforcement action, either referring the case to the Department of Justice for federal court action or proceeding to an administrative hearing to terminate federal funding to the school. Even in the absence of a complaint, DOJ and OCR have the authority to investigate colleges and universities suspected of failing to comply with Title VI (West-Faulcon, 2009; Department of Education, 2012; Department of Justice, 2001).

From an assessment point of view, the OCR burden-shifting approach is particularly appealing; it takes the formalistic framework of the disparate impact approach—an approach that relies on statistical evidence—and extends it by interrogating how we might

achieve educational goals through alternative means with less of a discriminatory effect on the disproportionately affected racialized group. This socio-contextual view of assessment is powerful as its interrogates how local decisions about test score interpretation can be put in conversation with larger social goals toward fairness and OTL. In the following example, we illustrate the benefits of the OCR burden-shifting approach to disparate impact in a writing assessment case while also detailing its limitations.

2.1.3 A final note about disparate impact today. Before continuing to an illustration of disparate impact analysis, it is important to note the recent Supreme Court decision handed down in June 2015. Much to the surprise of critics, the Court, again, upheld the viability of disparate impact theory in *Texas Department of Housing and Community Affairs v. The Inclusive Communities Project*. In its decision regarding disparate impact theory under the Fair Housing Act (FHA), however, the Court placed various restrictions on disparate impact claims. Writing the majority opinion, Justice Kennedy stated, "Recognition of disparate-impact liability under the FHA . . . plays a role in uncovering discriminatory intent: It permits plaintiffs to counteract unconscious prejudices and disguised animus that escape easy classification as disparate treatment" (*Texas Department of Housing and Community Affairs v. The Inclusive Communities Project*, 2015, p. 17). Yet, the Court also ruled that racial imbalance alone cannot substantiate disparate impact claims and that lower courts should

> examine with care whether a plaintiff has made out a prima facie case of disparate impact[,] and prompt resolution of these cases is important. A plaintiff who fails to allege facts at the pleading stage or produce statistical evidence demonstrating a causal connection cannot make out a prima facie case of disparate impact. (*Texas Department of Housing and Community Affairs v. The Inclusive Communities Project*, 2015, p. 21)

Among other limitations, the Court also ruled that "even when courts do find liability under a disparate-impact theory, their remedial orders must be consistent with the Constitution," "should concentrate on the elimination of the offending practice that 'arbitrar[ily] . . . operate[s] invidiously to discriminate on the basis of rac[e]'" and "should strive to design them to eliminate racial disparities through race-neutral means" (p. 22).

In the end, despite critics' predictions that disparate impact would be struck down by the current Supreme Court, the precedent remains in place. Nevertheless, methodological connections between statistical data, consequence, and remedy remain in flux. This trajectory from statistical evidence to consequence to remedy is a powerful, distinct approach for advancing fairness—an approach that we demonstrate in the remainder of this article, using the burden-shifting approach outlined by the U.S. Department of Education.

3.0 Demonstration of the OCR Approach in a Writing Assessment Case

In previous work (Poe, Elliot, Cogan, & Nurudeen, 2014), we demonstrated the application of the OCR burden-shifting approach in a writing program. As we argued in our case drawn from an institutional dataset at a college we called Brick City University, the disparate impact approach is a valuable tool for self-study and is particularly relevant in the use of writing program assessment data, such as placement exams, portfolio assessment, and other kinds of proficiency testing.

Brick City University is a public four-year, doctorate-granting institution in Newark, New Jersey. Brick City has an acceptance rate of 65% and most students come to Brick City with a 3.1-3.5 high school GPA. Demographic percentages and SAT score comparisons are shown in Table 3 (College Board, *State Profile Report: New Jersey*, 2012; College Board, *Total Group,* 2013).

Table 3: Descriptive Statistics for Brick City University Admitted Students (n = 844)

		Mean SAT Writing Scores			Writing Placement		Graduation
Group	**Number and Percent**	**Brick City Admitted**	**New Jersey**	**National**	**Basic Writing Number and Percent**	**First Year Writing Number and Percent**	**Within six years of admission**
Overall	N/A	519 (*SD* = 84)	499 (*SD* = 116)	488 (*SD* = 114)	173 (24%)	671 (76%)	49%
African American	107 (13%)	493 (*SD* = 68)	417 (*SD* = 97)	417 (*SD* = 94)	50 (47%)	57 (53%)	40%
Native American	9 (1%)	504 (*SD* = 76)	458 (*SD* = 112)	462 (*SD* = 103)	2 (22%)	7 (78%)	47%
Asian	191 (23%)	526 (*SD* = 92)	566 (*SD* = 131)	528 (*SD* = 129)	29 (15%)	162 (85%)	59%
Hispanic	200 (24%)	491 (*SD* = 76)	440 (*SD* = 102)	443 (*SD* = 92)	57 (28%)	143 (72%)	47%
White	337 (39%)	538 (*SD* = 83)	522 (*SD* = 103)	515 (*SD* = 103)	35 (10%)	302 (90%)	54%

As Table 3 shows, African American, Native American, Hispanic, and white students admitted to Brick City have higher SAT scores than both the state and national averages. Asian students have slightly lower scores. However, through the writing placement exam—a locally developed timed, impromptu exam (see Poe, Elliot, Cogan, & Nurudeen, 2014 for more information)—47% of African American students, 22% of Native American students, 28% of Hispanic students, 10% of white students, and 15% of Asian students place into basic writing. Regarding six-year graduation rates of all students, 59% of Asian and 54% of white students at Brick City University graduate within six years. Only about 40% of African American students and about 47% of Hispanic and Native American students graduate within six years.

Since Brick City graduation rates are similarly low for all students placed in basic writing,[11] the fairness issue for Brick City was not whether some students were required to take basic writing, rather whether that requirement was doing harm to some groups more than others. Let us emphasize here that differences in test scores alone do not constitute disparate impact; students come to college with different writing proficiencies. Rather, disparate impact occurs when a facially-neutral test places an unfair disadvantage on one group versus another. In the Brick City case, the test meant that certain groups of students were placed into a course—basic writing—that seemed to have a disproportionately negative effect on those students' educational outcomes, i.e., graduation rates.[12] If the students placed into basic writing were graduating at the same rate as other students, it would be difficult to show disparate impact because the course would seem to have no effect on educational outcomes. The question at Brick City, thus, was whether the high remediation rates for African American and Hispanic students into basic writing might be causing a disproportionate impact on those students' graduation rates.

To conduct their fairness assessment of the consequences of basic writing, the Brick City writing program provided a three-phase inquiry using the OCR burden-shifting approach.

3.1 Step 1—Does the School District Have a Facially Neutral Policy or Practice that Produces an Adverse Impact on Students of a Particular Race, Color, or National Origin When Compared to Other Students?

Using the placement exam data, we applied the four-fifths rule. As shown in Table 4, using white students as the benchmark group, the four-fifths rule was not violated for Asian, Native American, or Hispanic students. The rule, however, was violated for African American students.

Table 4: Four-fifths Analysis of Brick City University's Writing Placement Results

	Total Students	White Students	Asian Students	Hispanic Students	Native American	African American
Total Population	844	337	191	200	9	107
Number of Students in Group Tracked to First Year Writing	671	302	162	143	7	57
Percent of Students in Group Tracked to First Year Writing	80%	90%	85%	72%	78%	53%
Four-Fifths Threshold *(.8 x Percentage of White Students Tracked to First Year Writing)*	72% (.8*.9=.72)					
Four-Fifths Rule Violated?	--	**N/A**	**No**	**No**	**No**	**Yes**

3.2 Step 2—Can the School District Demonstrate that the Policy or Practice is Necessary to Meet an Important Educational Goal?

After statistical analysis revealed that Brick City placement testing had an adverse impact on African American students, Brick City would then need to articulate how the placement exam supports an educational goal. This empirical inquiry could include evaluating whether the construct representation of writing that the placement exam measures is accurate for college-level writing; ensuring that the placement exam assesses those traits that are most likely to result in difficulties in college-level writing; documenting that the basic writing curriculum addresses those traits; and demonstrating that the placement exam is significantly correlated with students' performance in subsequent first-year writing courses. Note here that the writing program may not be able to identify the impact of basic writing on graduation rates, but it can make a connection between remediation and persistence into first-year courses, which has been shown to be predictive of continued success in college (Complete College America, 2015b).

3.3 Step 3—Are there Comparably Effective Alternative Policies or Practices that Would Meet the School District's Stated Educational Goal with Less of a Discriminatory Effect on the Disproportionately Affected Racial Group; or, is the Identified Justification a Pretext for Discrimination? (Department of Education, 2014, p. 8)

In the final step of the OCR burden-shifting approach, Brick City would then explore alternatives available that met the school's stated educational goal with less of a burden on African American students. At this stage, the discourse and processes of assessment change dramatically. Rather than looking solely to test scores, this final phase of the OCR method invites stakeholders to participate in curricular reform while maintaining the educational goals for writing instruction. In the Brick City case, a corequisite option was selected.

In making this selection, Brick City test designers followed the guidance of Standard 3.20:

When a construct can be measured in different ways that are equal in their degree of construct representation and validity (including freedom from construct-irrelevant variance), test users should consider, among other factors, evidence of subgroup differences in mean scores or percentages of examinees whose scores exceed the cut scores, in deciding which test and/or cut scores to use. (AERA, APA, & NCME, 2014, p. 72)

4.0 Group Classification Considerations Using the Disparate Impact Approach

The Brick City case provides much optimism; it relies on established empirical methods for evaluating disparate impact, demands the articulation of curricular goals, and invites curriculum innovation while maintaining consistent educational goals. Yet, the burden-shifting approach is not without problems. Legal critics have argued, for example, that disparate impact analysis is reactive rather than proactive, thus making it out-of-step with international human rights standards (Hunter & Shoben, 2014), that there are not comparable methods or standards for evaluating intentional discrimination (Selmi, 2006; Willborn, 1985), and that the statistical measures suggestive of adverse impact, such as the *Z* value greater than 1.96 standard deviations and four-fifths rule are arbitrary.

From a measurement perspective, the burden-shifting approach has another challenge—strength of sampling plan. Strength of sampling plan is a problem that has long vexed the measurement communities, especially with regard to small populations (Kane, 1982, 2011; Linn, 1989). Thus, when Standard 3.2 makes the seemingly straightforward recommendation that "those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test," researchers should take to heart that this is not a straightforward process (AERA, APA, & NCME, 2014, p. 64). Because of the challenges of statistical analysis using small populations, new techniques such as resampling (Yu, 2003), including Monte Carlo sampling (Yu, 2003), have been tools to ensure robust group sizes for statistical analysis even for small populations. This, however, is not the case in local writing assessment, where resampling may not be a viable technique for reporting purposes or data on group performances may not be collected. Ultimately, many writing program administrators are faced with the reality of having insufficiently large sample sizes from which to conduct comparative group analysis.[13]

Adoption of the burden-shifting approach, thus, requires some caution, as the issues identified in the legal and measurement literature are worthy of further discussion. For our purposes here—and to keep this article relatively brief—we want to address one concern that has been overlooked by both legal and assessment scholars—characterization of groups. By characterization of groups we mean how group populations are defined and identified. As we think of it, there are three questions that can be used to guide this inquiry: (1) Do the group identifications describe meaningful traits for the group that encompass social equity concerns? (2) Are the inferences drawn from the group identifications sufficiently grounded in the contextual conditions for that group? And (3) Are there combinations of variables that suggest different inferences are salient for focal groups? To answer these questions, we discuss three issues: group specification, demographic shifts, and intersectionality.

It should be noted that in the following discussion, we use the term "group identity" here rather than subgroup, as it more accurately reflects today's demographic realities; subgroup may be statistically useful but socially demeaning. There are no longer groups and subgroups, simply groups.

4.1 Group Specification

The issue for population specification in regard to fairness is two-fold. First, we cannot assume that the group specification today is without its flaws—a complexity that is evidenced nationally in the shifting categories used on census records and in legal decisions (Lopez, 1997). Second, there must be a commitment to ensuring that the criteria used to define groups is meaningful across groups in order to provide the kinds of evidence needed to make fairness claims. Thus, population specification—"the ways in which cultural groups are defined and, therefore, the criteria used to determine when an individual belongs to a certain cultural group" (Basterra, Trumbull, & Solano-Flores, 2011, p. 9)—is complex and should be treated as such.

A lesson from history is instructive here: In Cleary's classic 1966 study in which she argued a test is biased "if too high or too low a criterion score is consistently predicted for members of the subgroup when the common regression line is used" (p. 1), she engaged with two problems of sampling: group identification and sample size. First, as she explained, "the scarcity of Negro students in the integrated colleges is disturbing," thus leaving her a small number of schools for the study (p. 5). In regard to group identification, she explained, "Most schools had no record of the race of their individual students" (p. 5). In such cases, Cleary relied on the

judgments of two "persons" who "examine[d] independently the standard identification pictures in the school files" (p. 6). Based upon the judges' assessment, the students were assigned a racialized identity. In instances where the judges could not agree, "the student was classified as white" (p. 6). She used NAACP records to corroborate judges' ratings.

As cultural critics would expect, Cleary's "look test" method of racial identification—a method also used by Pfeifer & Sedlacek (1971)—was less than perfect, a point she acknowledged in noting how she addressed outliers:

> five students not on the NAACP list had been classified as Negro, and one student on the NAACP list had been classified as white. The five students not on the NAACP list were retained as Negroes after further examination of the identification pictures. The race code of the one student who was on the NAACP list but who had not been classified as Negro was changed to Negro. (p. 6)

In other words, students who "looked black" stayed in the Negro sample for the study, and the student who phenotypically passed for white but was listed on the NAACP record was replaced into the Negro sample. We would argue that such methodological choices are reflective of U.S. historical norms regarding the one-drop rule, not scientific method.

Today, most studies such as our Brick City example rely on student's self-reported racialized identity using the Office of Management and Budget categories (Office of Management and Budget, 1995).[14] While racialized identity may be tied to federal census categories set by the Office of Management and Budget, other group identifications such as socioeconomic status and linguistic identity are even more complicated. Family income and educational levels, for example, have conventionally been used as proxies for socioeconomics status (SES) (ACT, 2014; Sacket et al., 2009). In K-12 studies, researchers may also use qualification for free or reduced meals as an indicator of socioeconomic status.[15] And the National Center for Educational Statistics (2012) has recommended:

> Family income and other indicators of home possessions and resources, parental educational attainment, and parental occupational status should be considered components of a core SES measure . . . Neighborhood and school SES could be used to construct an expanded SES measure. (2012, p. 5)

Among the many criticisms of self-economic status indicators as meaningful markers of group identity are that income (e.g, annual salary) and education are not what separate racialized groups. Instead, it is wealth (e.g., investments, home ownership, etc.). In 2006, the median net worth of a white family was $120,900; for people of color, it was $17,100 (Liu et al., 2006, p. 3). In 2009, the median wealth of white families was $113,149; for Latino families it was $6,325 and for black families it was $5,677 (Kochar, Fry, & Taylor, 2011). In a study conducted by the Institute for Assets and Social Policy at Brandeis University, researchers traced the same households over 25 years. During that time, the total wealth gap between white and African-American families increased from $85,000 in 1984 to $236,500 in 2009 (Shapiro, Meschede, & Osoro, 2013, p. 1). Home ownership, income, unemployment, education, and inheritance were the main drivers of wealth inequality with home ownership being the largest predictor for wealth gap (p. 3). Moreover, additional income gains, inheritance, other financial supports, and marriage yielded different rates of return—for example, a $1 increase in income for a white family converted to $5.19 of wealth. That same dollar increase for Black families yielded 69 cents in wealth. Finally, a criticism of conventional socioeconomic indicators is that they do not reflect historical legacy. Rubin et al. (2014) have argued that SES is different than social class with SES referring to one's current social and economic situation and social class referring to one's sociocultural background (p. 196). SES, they argued, may be quite variable while social class tends to remain more fixed.

Linguistic identity is also illustrative here. Much has been written in the field of second language writing regarding how researchers might best capture the nuances of contemporary multilingual identity (e.g., Shohamy, 2011). This attention to evolving definitions of World Englishes and "languaging" is often not found in the assessment literature (Dryer, 2016). For example, in a recent study conducted by Sinharay, Dorans, and Liang (2011) regarding fairness procedures for test-takers whose first language is not English, they used a rather thin definition to determine group specification:

> For illustrative purposes, we use the first-language status of a test taker as a surrogate for language proficiency and describe an approach to examining how the results of fairness procedures are affected by inclusion or exclusion of those who report that English is not their first language in the fairness analyses. (p. 25)

4.2 Shifting Demographics

Demographic shifts are the largest challenges to making longitudinal claims about fairness for two reasons (Aud, Fox, & KewalRamani, 2010). First, traditional categories used to describe racial/ethnic groups may belie fundamental changes within those groups. Second, the use of white students as the reference group may no longer be appropriate if they are no longer the majority population—or even the population that reports back the highest scores on tests and other assessments. In such cases where group identification is shifting, researchers must proceed with extra caution in making inferences. This point cannot be understated.

If the gold standard of validity is to be prediction, then longitudinal claims must be interrogated carefully.

Again, history is illustrative here: The 1966 Equality of Educational Opportunity report, also known as the Coleman report for its lead author, sociologist James Coleman, was submitted in response to the Civil Rights Act of 1964, which ordered a survey and a report to the President and the Congress

> concerning the lack of availability of equal educational opportunities for individuals by race, color, religion, or national origin in educational institutions at all levels in the United States, in territories and possessions, and the District of Columbia. (§402)

The Coleman report researchers were tasked with determining the extent of racial segregation in U.S. schools and "whether the schools offer equal educational opportunities in terms of a number of other criteria which are regarded as good indicators of educational quality" (Coleman et al., 1966, p. iii). The researchers did not review differences by religion or country of origin, and instead relied on six racial categories: Negroes, American Indians, Oriental Americans, Puerto Ricans living in the continental United States, Mexican Americans, whites other than Mexican Americans and Puerto Ricans (Coleman et al., 1966, p. iii).

With respect to demographics, there are two lessons from the Coleman report. First, the racial designations used by the Coleman researchers 50 years ago are out-of-sync with today's terminology. Moreover, in contrast to the Civil Rights era, today most immigrants are classified under Asian or Hispanic group designations (Migration Policy Institute; Census, 1999).[16] As previously demonstrated (Inoue & Poe, 2012), longitudinal claims about group performances can lead to inaccurate conclusions when group ethnic formations are not compared. In the Inoue & Poe study, results of the California State University English Placement Test were traced over 25 years, noting that the results suggested a decline in the performance Asian students. Upon closer investigation, it was determined that the ethnic groups that comprised the Asian group had shifted dramatically during the time period under study. While previously students had been of Chinese background, more recent students were Hmong, a group that has strong agrarian ties and, given their refugee status across multiple countries, often does not have a history of formal education within families.

Second, the architects of the Coleman report—following the history of U.S. legal and social precedent—constructed a narrative of the U.S. that is based on distinct racialized categories and north/south geographic comparisons. White students were always the demographic group to which African American students were compared. In today's shifting U.S. demographics—a demographic change that has been called "stunning" (Teixeira, Frey, & Griffin, 2015, p. 2)—white students may no longer be the appropriate reference group, thus shifting the entire referential frame by which group comparisons are made.

In Brick City's case, Hispanic students now make up the second largest group of admitted students (200 Hispanic students versus 337 white students). In the last 10 years, while the number of Asian American students and African American students has remained consistent, the number of Hispanic students admitted to Brick City has doubled, reflecting the changing demographic patterns of its regionally-serving identity. If such a trend continues, within the next decade Hispanic students will become the reference group against which all others will be compared.

4.3 Intersectionality

Likely the most methodologically challenging aspect of disparate impact analysis is intersectionality—the multidimensionality of identity that reveals intergroup differences. Crenshaw's scholarship in legal journals (1989, 1991) is widely cited on intersectionality. For Crenshaw, discrimination challenges often are imbued with a flawed logic that separates race from gender: "…in race discrimination cases, discrimination tends to be viewed in terms of sex- or class-privileged Blacks; in sex discrimination cases, the focus is on race- and class- privileged women" (1989, p. 140). As a result, those who are "multiply-burdened" are marginalized and claims are obscured "that cannot be understood as resulting from discrete sources of discrimination" (Crenshaw, 1989, p. 140). What Crenshaw posits then is, for example, that the effects of race/gender/class are more subtle and perhaps greater than race plus gender plus class. Further disaggregating columns in a spreadsheet or conducting a multiple regression analysis will not reveal the cascading effects of a legacy of brutality. Interestingly, it is a subtlety that Johnson, too, pointed out in his 1965 commencement speech at Howard University:

> For Negro poverty is not white poverty. Many of its causes and many of its cures are the same. But there are differences--deep, corrosive, obstinate differences--radiating painful roots into the community, and into the family, and the nature of the individual. These differences are not racial differences. They are solely and simply the consequence of ancient brutality, past injustice, and present prejudice. (Johnson, 1965)

In reviewing decades of assessment literature, it is striking how traditionally few researchers looked at combinations of variables. Today, researchers like Zwick & Green (2007) and Zwick & Himelfarb (2011) provide some useful direction in that they revisit existing wisdom about prediction of SAT scores and high school grades through the lens of school resources and within versus across school comparisons. Yet, more is to be done in the development of fairness methods. Specifically, further advancement is

needed to understand the cascading effects of multiple variables as well as to understand intergroup differences (e.g., If we start out to look at differences between Asian and White students, we are likely to find them without attending to the differences within the performance of Asian students.). These “indices of fairness that reflect more complex matrices,” as Petersen and Novick called them 40 years ago, should look at identity clusters within groups (e.g., African American women from middle class backgrounds) to help researchers make more nuanced claims about fairness and ensure that researchers do not assume homogeneity within groups. Bottom line: Without nuance, meaningful change is unlikely.

At the 1965 Howard University commencement address, Lyndon B. Johnson declared:

> You do not take a person who, for years, has been hobbled by chains and liberate him, bring him up to the starting line of a race and then say, “you are free to compete with all the others,” and still justly believe that you have been completely fair. Thus it is not enough just to open the gates of opportunity. All our citizens must have the ability to walk through those gates.

Our goal in writing this article was to advance disparate impact theory, and more specifically the burden-shifting approach, as a conceptual and methodological framework for fairness. Like Kane, we are inclined to define fairness and validity broadly (2010), but our fear is that collapsing fairness into validity will result in the inattention to fairness. As we have suggested, methodological advancements such as those by Zwick and Green (2007), Zwick and Himelfarb (2011), and Xi (2010) are useful, important, and insufficient if they are not viewed as part of the process of developing a rigorous conceptual and methodological framework for fairness. Such a framework must include questions of access, response processes, test score interpretation, and social consequence.

Disparate impact theory and the burden-shifting approach as outlined by the OCR provides a theory and a method by which we can recognize that past inequality has consequences today. The approach combines empiricism and contextualization—i.e., data do not speak to themselves without the force of history and social action. In doing so, the OCR process invites reflection; it encourages us to think expansively, beyond comfortable, known strictures. Finally, the disparate impact approach has been sustainable, weathering the political shifts of the Supreme Court and the shifting social and demographic changes of the U.S. over the last 50 years.

Of course, there remain questions. Disparate impact analysis, for example, has not been evaluated using intersectional identities: Are the effects of unintentional discrimination different for African American women, for example, than for African Americans as a group? Likewise, under what time scales can disparate impact analysis be meaningful when racialized group identifications can shift dramatically in a few generations? And without interrogation of group identification during step 1 of the disparate impact analysis, arguments made about fairness could be made only of gossamer. Such questions should not arouse suspicions about the viability of disparate impact. Instead, through the pursuit of these questions and others, disparate impact theory and the burden-shifting approach can be enriched and deepened for the purpose of fairness studies.

In the end, if equitability is to be valued, it must be seen. Fairness in theory cannot be an afterthought to validity or reliability. Fairness in action demands local attention in which we repeatedly question how we can achieve equitable results with less adverse impact—in which “the rights of every man are diminished when the rights of one man are threatened” (Kennedy, 1963). Test scores may reflect social inequality, but the *use* of test scores works to create that social inequality. Racial isolation and structural inequality are not merely reflective of such social mechanisms; social mechanisms work to sustain invisibility, racialized isolation, and structural inequality. The creation of opportunity structures through approaches such as disparate impact analysis holds the potential to provide visibility, community, and equity.

Acknowledgements

We would like to thank our colleagues Bob Broad, Ellen Cushman, and David Slomp as well as two anonymous *JWA* reviewers for their comments on this article. We would also like to thank *JWA* editors, Diane Kelly-Riley and Carl Whithaus, for their support of this special issue. A special thanks goes to Norbert Elliot for being a saintly respondent and good friend.

References

34 C.F.R. §100.3 (2015).

42 U.S.C. § 1983 (2012).

42 U.S.C. § 2000d (2012).

42 U.S.C. § 2000d-1 (2012).

740 Ill. Comp. Stat. § 23/5 (2004).

Abernathy, C. (1981). Title VI and the constitution: A regulatory model for defining "discrimination." *Georgetown Law Journal, 70,* 1–49.

Abernathy, C. (2006). Legal realism and the failure of the "effects" test for discrimination. *Georgetown Law Journal, 94*(2), 267–319.

Center for Culturally Responsive Evaluation and Assessment. (n.d.) About CREA. Retrieved from http://education.illinois.edu/CREA/about-crea

ACT. (2014). *The condition of college & career readiness 2014: Students from low-income families*. Retrieved from www.act.org/readiness/2014

AERA, APA, & NCME. (1974). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

APA Task Force on Employment Testing of Minority groups. (1969). Job testing and the disadvantaged, *American Psychologist, 24*, 637–50.

Adams, P., Gearhart, S., Miller, R. & Roberts, A. (2009). The accelerated learning program: Throwing open the gates. *Journal of Basic Writing, 28*(2), 50–69.

Alexander v. Sandoval, 532 U.S. 275 (2001).

Anderson, T., Scum, D., & Twining, W. (2005). *Analysis of evidence* (2nd ed.). Cambridge, UK: Cambridge University Press.

Aronson, J., Fried, C., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, 38(2), 113–125.

Aud, S., Fox, M., & KewalRamani, A. (2010). *Status and Trends in the Education of Racial and Ethnic Groups*. National Center for Educational Statistics Report (NCES 2010-015). Washington, DC: Department of Education.

Basterra, M. R., Trumbull, E., & Solano-Flores, G. (Eds.). (2011). *Cultural validity in assessment: Addressing linguistic and cultural diversity*. New York: Routledge.

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

Bennett, R., Kane, M., & Bridgeman, B. (2011). Theory of action and validity argument in the content of through-course summative assessment. Educational Testing Service.

Black, D. (2002). Picking up the pieces after *Alexander v. Sandoval*: Resurrecting a private cause of action for disparate impact. *North Carolina Central Law Review, 81*(1), 56–391.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge, UK: Cambridge University Press.

Braceras, J. (2002). Killing the messenger: The misuse of disparate impact theory to challenge high-stakes educational tests. *Vanderbilt Law Review, 55*, 1111-1203.

Breland, H., & Ironson, G. (1976). DeFunis Reconsidered: A Comparative Analysis of Alternative Admissions Strategies. *Journal of Educational Measurement, 13*(1), 89–99.

Cal. Gov't Code § 11135 (West Supp. 2015).

Cal. Gov't Code § 11139 (West Supp. 2015).

Camilli, G. (2006). Test Fairness. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 221–256). Westport, CT: Rowman & Littlefield Publishers.

Castañeda v. Partida, 430 U.S. 482 (1977).

Census Bureau (n.d.). Hispanic origin. Retrieved from http://www.census.gov/topics/population/hispanic-origin.html

Census Bureau (1999). Race and Hispanic origin of the population by nativity: 1850 to 1990. Retrieved from http://www.census.gov/population/www/documentation/twps0029/tab08.html

Chapelle, C., Enright, M. & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13.

Civil Rights Act, 42 U.S.C. § 2000e (2012).

Cleary, T. (1966). *Test bias: Validity of the scholastic aptitude test for Negro and white students in integrated colleges*. (College Entrance Examination Board RDR-65-6, No. 18). Princeton, NJ: Educational Testing Service.

Cohen, G., Steele, C., Ross, L. (1999). The mentor's dilemma: Providing critical feedback across the racial divide. *Personality and Social Psychology Bulletin, 25*(10), 1302–1318.

Cole, N. & Zieky, M. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*(4), 369–382.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of educational opportunity*. Washington, DC: National Center for Educational Statistics.

College Board. (2012). *State Profile Report: New Jersey*. New York: College Board.

College Board. (2013). *Total Group Profile Report*. New York: College Board.

Complete College America. (2011). *Time is the enemy: The surprising truth about why today's college students aren't graduating ... and what needs to change*. Washington, DC: Complete College America.

Complete College America. (2012). *Remediation: Higher education's bridge to nowhere*. Washington, DC: Complete College America.

Complete College America. (2015a). The results are in. Corequisite remediation works. Retrieved from http://completecollege.org/the-results-are-in-corequisite-remediation-works

Complete College America. (2015b). *Corequisite Remediation: Spanning the Completion Divide*. Washington, DC: Complete College America. Retrieved from http://completecollege.org/spanningthedivide/wp-content/uploads/2016/01/CCA-SpanningTheDivide-ExecutiveSummary.pdf

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum, 140*, 139–167.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review, 43*(6), 1241–1299.

Cronbach, L. (1976). Equity in selection: Where psychometrics and political philosophy meet. *Journal of Educational Measurement, 13*(1), 31–42.

Daly, A. (2006). How to speak American: In search of the real meaning of "meaningful access" to government services for language minorities. *Penn State Law Review, 110*(4), 1005–1046.

Davis, E. (2006). Unhappy parents of limited English proficiency students: What can they really do. *Journal of Law and Education,*

35(2), 277–288.

Child Nutrition Programs–Income Eligibility Guidelines, 80(61) Fed. Reg. 17026-17027 (2015). Retrieved from http://www.gpo.gov/fdsys/pkg/FR-2015-03-31/pdf/2015-07358.pdf

Department of Education. (2012, December 5). *Compliance review letter metropolitan school district of Pike Township* [Letter to Nathaniel Jones]. Office for Civil Rights, Washington DC. Retrieved from https://www2.ed.gov/about/offices/list/ocr/docs/investigations/05085002-a.pdf

Department of Education. (2014). *Dear colleague letter: Resource compatibility* [Letter on educational opportunity]. Office for Civil Rights, Washington DC. Retrieved from http://www2.ed.gov/about/offices/list/ocr/letters/colleague-resourcecomp-201410.pdf

Implementation of the Fair Housing Act's Discriminatory Effects, 78(32) Fed. Reg. 11460-11482 (2013). Retrieved from http://portal.hud.gov/hudportal/documents/huddoc?id=discriminatoryeffectrule.pdf

Department of Justice. (2002). *Guidance to federal financial assistance recipients regarding Title VI prohibition against national origin discrimination affecting limited English proficient persons*. Washington DC: U.S. Department of Justice. Retrieved from http://www.hhs.gov/ocr/civilrights/resources/laws/revisedlep.html

Dryer, D. (2016). Appraising translingualism. *College English, 78*(3), 274–283.

Elliot, N., Deess, P., Rudniy, A., & Joshi, K. (2012). Placement of students into first-year writing courses. *Research in the Teaching of English, 46*(3), 285–313.

Elul, H. (1998). Making the grade, public education reform: The use of standardized testing to retain students and deny diplomas. *Columbia Human Rights Law Review, 30*, 495–536.

Every Child Succeeds Act, S. 1177 (2015). Retrieved from https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf

Fassold, M. (2000). Disparate impact analyses of TAAS scores and school quality. *Hispanic Journal of Behavioral Sciences, 22*(4), 460–480.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement, 39*(2), 133–147.

Ladson-Billings, G. (1994). *The dreamkeepers*. San Francisco: Jossey-Bass Publishing Co.

Glau, G. (1996). The "stretch program": Arizona State University's new model of university-level basic writing instruction. *WPA: Writing Program Administration*, 20(1/2), 79–91.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology, 24*(6), 645–662.

Green III, P. (1996). Can Title VI prevent law schools from adopting admissions practices that discriminate against African-Americans? *Southern University Law Review, 24*(2), 237–261.

Grego, R., & Thompson, N. (1995). The writing studio program: Reconfiguring basic writing/freshman composition. *WPA: Writing Program Administration, 19*(1/2), 66–79.

Grego, R., & Thompson, N. (2007). *Teaching/Writing in thirdspaces: The studio approach.* Carbondale, IL: Southern Illinois University Press.

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Haertle, E., Moss, P., Pullin, D., & Gee, J. (2008). Introduction. In P. Moss, D. Pullin, J. Gee, E. Haertle, & L. Young (Eds.) *Assessment, equity, and opportunity to learn* (pp. 1-16). Cambridge UK: Cambridge UP.

H.R. Misc. Doc. No. 124, 88th Cong., 1st Sess. 3, 12 (1963).

Hunter, R., & Shoben, E. (2014). Disparate impact discrimination: American oddity or internationally accepted concept. *Berkeley Journal of Employment & Labor Law, 19*(1), 108–152.

Hussar, W. J., & Bailey, T. M. (2011). *Projections of education statistics to 2020* (NCES 2011-026). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Inoue, A. & Poe, M. (2012). Racial formations in two writing assessments: Revisiting White and Thomas' findings on the English Placement Test after 30 Years. In N. Elliot & L. Perelman (Eds.) *Writing assessment in the 21st Century: Essays in honor of Edward M. White* (pp. 343-36). New York, NY: Hampton Press.

Isaacs, E. (forthcoming). *Writing at the State U: Writing instruction and writing program administration at 106 U.S. representative institutions.*

Jaeger, R. (1976). A word about the issue. *Journal of Educational Measurement, 13*(1),1.

Johnson, O. (2014). The agency roots of disparate impact. *Harvard Civil Rights-Civil Liberties Law Review, 49*, 125–154.

Johnson, L. (1965, June 4). To fulfill these rights. Commencement Address at Howard University, Washington, DC.

Kane, M. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*(2), 125–160.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64 Westport, CT: Rowman & Littlefield Publishers.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177–182.

Kane, M. (2011). The error of our ways. *Journal of Educational Measurement, 48*(1), 12–30.

Kane, M. (2012). *Validity, fairness, and testing*. Presentation at the Conference on Conversations on Validity Around the World, Teachers College, New York, NY. https://www.ets.org/c/18486/pdf/19633_ets_tc_validityconference_Kane%202012_03_28.pdf

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

Kennedy, J. (1963, June 11). A report to the American people on civil rights. [Radio and television address]. New York, NY: Columbia Broadcasting System.

Kidder, W., & Rosner, J. (2002). How the SAT creates built-in-headwinds: An educational and legal analysis of disparate impact. *Santa Clara Law Review, 43*(1), 131–211.

Kochhar, R., Fry, R., & Taylor, P. (2011). *Wealth gaps rise to record highs between Whites, Blacks, Hispanics*. Washington, D.C.: Pew Research Center.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge, UK: Cambridge University Press.

Langenfeld, T. (1997). Test fairness: Internal and external investigations of gender bias in mathematics testing. *Educational Measurement: Issues and Practice, 16*(1), 20–26.

Lindsay, P., & Crusan, D. (2011, December 21). How faculty attitudes and expectations toward student nationality affect writing assessment. *Across the Disciplines, 8*(4). Retrieved from http://wac.colostate.edu/atd/ell/lindsey-crusan.cfm

Linn, R. (1973). Fair test use in selection. *Review of Educational Research, 43*, 139–161.

Linn, R. (1976). In search of fair selection procedures. *Journal of Educational Measurement, 13*(1), 53–58.

Linn, R. (Ed.). (1989). *Intelligence, theory, and public policy*. Urbana, IL: University of Illinois.

Lioi, A. & Merola, N. (2012). The muse of difference: Race and writing placement at two elite art schools. In A. Inoue & M. Poe (Eds.), *Race and writing assessment* (pp. 155–168). New York: Peter Lang.

Liu, M., Robles, B., Leonder-Wright, B., Brewer, R., & Adamson, R. (2006). *The color of wealth: The story behind the U.S. racial wealth divide.* New York: The Fair Press.

Lopez, I. (1997). *White by law: The legal construction of race*. New York City: NYU Press

Lu, C. & Suen, H. (1995). Assessment approaches and cognitive style. *Journal of Educational Measurement, 32*(1), 1–17.

Marefat, F., & Heydari, M. (2016). Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. *Assessing Writing, 27*, 24–36.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012–1027.

Messick. S. (1989a) Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Migration Policy Institute. (n.d.). Retrieved from http://www.migrationpolicy.org/programs/data-hub/charts/largest-immigrant-groups-over-time

Mislevy, R.J., Haertel, G., Cheng, B., Ructtinger, L., DeBarger, A., Murray, E., …Vendlinski. T. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *19*, 121–140.

Moss, P., Pullin, D., Gee, J., Haertle, E., & Young, L. (Eds.). (2008). *Assessment, equity, and opportunity to learn*. Cambridge UK: Cambridge UP.

Murray, V. (2008). *The high price of failure in California: How inadequate education costs schools, students, and society.* San Francisco: Pacific Research Institute.

National Center for Educational Statistics. (2012). *Improving the measurement of socioeconomic status for the National Assessment of Educational Progress: A theoretical foundation.* Retrieved from https://nces.ed.gov/nationsreportcard/pdf/researchcenter/Socioeconomic_Factors.pdf

Nieto, S. (2013). *Finding joy in teaching students of diverse backgrounds: Culturally responsive and socially just practices in U.S. classrooms*. Portsmouth, NH: Heinneman.

Nguyen, H., & Ryan, A. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93*(6), 1314–1334.

Office of Management and Budget. (1995). *Standards for maintaining, collecting, and presenting federal data on race and ethnicity*. Washington, DC: Federal Register. Retrieved from https://www.whitehouse.gov/omb/fedreg_race-ethnicity

Or. Rev. Stat. § 659.850 (West Supp. 2015).

Perry, P. (1991). Two faces of disparate impact discrimination. *Fordham Law Review, 59*, 523–595.

Petersen, N., & Novick, M. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13*(1), 3–29.

Pfeifer, C., & Sedlacek, W. (1971). The validity of academic predictors for black and white students at a predominantly white university. *Journal of Educational Measurement, 8*(4), 253–261.

Phillips, S., & Camara, W. (2006). Legal and ethical issues. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 733–755). Westport, CT: Rowman & Littlefield Publishers.

Poe, M., Elliot, N., Cogan, J., & Nurudeen, T. (2014). The legal and the local: Using disparate impact analysis to understand the consequences of writing assessment. *College Composition and Communication, 65*(4), 588–611.

Pollock, M. (2005). Keeping on keeping on: OCR and complaints of racial discrimination 50 years after Brown. *The Teachers College Record, 107*(9), 2106–2140.

Popham, J. (2012). *Assessment bias: How to banish it. Mastering assessment: A self-service system for educators* (2nd ed.). Boston: Pearson.

Powe v. Miles, 407 F.2d 73 (1968).

Pullin, D. (2008). Assessment, equity, and opportunity to learn. In P. Moss, D. Pullin, J. Gee, E. Haertle & L. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 333-351). Cambridge UK: Cambridge UP.

Pullin, D. (2013). Legal issues in the use of student test scores and value added models (VAM) to determine educational quality. *Education Policy Analysis Archives, 21*(6), 1–27.

Pullin, D. (2014). *Performance, value, and accountability: Public policy goals and legal implications of the use of performance assessments in the preparation and licensing of educators*. Washington, D.C.: Council of Chief State School Officers and Center for Assessment, Learning and Equality of Stanford University.

Rendell-Baker v. Kohn, 457 U.S. 830 (1982).

Ricci v. DeStephano, 557 U.S. 557 (2009).

Richardson v. Lamar County Board of Education, 729 F.Supp. 806 (1989).

Robinson v. Kansas, 295 F.3d 1183 (10th Cir. 2002).

Rubin, M., Denson, N., Kilpatrick, S., Matthews, K., Stehlik, T., & Zyngier, D. (2014). "I am working-class": Subjective self-definition as a missing measure of social class and socioeconomic status in higher education research. *Educational Researcher, 43*(4), 196–200.

Ryan, J. (2003). Race discrimination in education: A legal perspective. *Teachers College Record, 105*(6), 1087–1118.

Sackett, P., Kuncel, N., Arneson, J., Cooper, S., & Waters, S. (2009). *Socioeconomic status and the relationship between the SAT® and freshman GPA: An analysis of data from 41 colleges and universities.* New York: College Board.

Save Our Valley v. Sound Transit, 335 F.3d 932 (9th Cir. 2003).

Sawyer, R., Cole, N. & Cole, J. (1976). Utilities and the issue of fairness in a decision theoretic model for selection. *Journal of Educational Measurement, 13*(1), 59–76.

Selmi, M. (2006). Was the disparate impact theory a mistake? *UCLA Law Review, 53*(3), 701–782.

Shapiro, T., Meschede, T., & Osoro, S. (2013). *The roots of the widening racial wealth gap: Explaining the black-white economic divide*. Research and Policy Brief. Waltham, MA: Institute on Assets and Social Policy.

Shermis, M., Lottridge, S., & Mayfield, E. (2015). The impact of anonymization for automated essay scoring. *Journal of Educational Measurement, 52*(4), 419–436.

Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *The Modern Language Journal, 95*(iii), 418–429.

Sinharay, S., Dorans, N., & Liang, L. (2011). First language of test takers and fairness assessment procedures. *Educational*

Measurement: Issues and Practice, 30(2), 25–35.

Sireci, S., & Green, P. (2005). Legal and psychometric criteria for evaluating teacher certification tests. *Educational Measurement: Issues and Practice, 19*(1), 22–31.

Sireci, S., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualization of validity. *Educational Measurement: Issues and Practice, 25*(3), 27–34.

Slomp, D., Corrigan, J. & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating Canadian large-scale writing assessments. *Research in the Teaching of English, 48*(3), 276–302.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553–573.

Solano-Flores, G. (2002). Assessing the cultural validity of assessment practices: An introduction. In M. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 3–21). New York: Routledge.

Smith, L., & Hambleton, R. (1990). Content validity studies of licensing examinations. *Educational Measurement: Issues and Practice, 9*(4), 7–10.

Soliday, M. (2002). *The politics of remediation: Institutional and student needs in higher education.* Pittsburgh, PA: University of Pittsburgh Press.

South Camden Citizens in Action v. New Jersey Department of Environmental Protection, 274 F.3d 771 (3d Cir. 2001).

Steele, C., & Aronson, J. (1998). Stereotype threat and the test performance of academically successful African Americans. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 401–427). Washington DC: Brookings Institution.

Sternglass, M. (1997). *Time to know them: A longitudinal study of writing and learning at the college level.* Mahwah, NJ: Erlbaum.

Stricker, L. J. (2008). *The challenge of stereotype threat for the testing community* (ETS RM-08-12). Princeton, NJ: Educational Testing Service.

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

Teixeira, R., Frey, W. H., & Griffin, R. (2015). *States of change: The demographic evolution of the American electorate, 1974-2060*. Washington, DC: Center for American Progress, American Enterprise Institute, & Brookings Institution.

Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 576 U.S. ___, 135 S. Ct. 2507 (2015).

U.S. Const. amend XIV.

Verdun, V. (2005). The big disconnect between segregation and integration. *Negro Educational Review, 56*(1), 67–82.

Walker, M., & Bridgeman, B. (2008). *Stereotype threat spillover and SAT® scores*. College Board. Research Report No. 2008-2.

Walton, G. & Spenser, S. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science, 20*(2), 1132–1139.

Washington v. Davis, 426 U.S. 229 (1976).

Watson, S. (1990). Reinvigorating Title VI: Defending health care discrimination—It shouldn't be so easy. *Fordham Law Review, 58*(5), 939–978.

West-Faulcon, K. (2009). The river runs dry: When Title VI trumps state anti-affirmative action laws. *University of Pennsylvania Law Review, 157*(4), 1075–1106.

Willborn, S. (1985). The disparate impact model of discrimination: Theory and limits. *The American University Law Review, 34*, 799–837.

Wilson v. Collins, 517 F.3d 421 (6th Cir. 2008).

Wislon, J., Olinghouse, N., McCoach, D., Santangelo, T., & Andrada, G. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing, 27*, 11–23.

Worrell, F. (2014). Forty years of Cross' nigrescence theory: From stages to profiles, from African Americans to all Americans. In J. Sullivan & A. Esmail (Eds.), *African American identity: Racial and cultural dimensions of the Black experience* (pp. 3–28). Lanham, MD: Lexington Books.

Yaeger, D., & Walton, G. (2011). Social-Psychological interventions in education: They're not magic. *Review of Educational Research, 81*(2), 267–301.

Yu, C. H. (2003). Resampling methods: concepts, applications, and justification. *Practical Assessment, Research & Evaluation, 8*(19). Retrieved from http://PAREonline.net/getvn.asp?v=8&n=19

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

Zwick, R., & Green, J. (2007). New perspectives on the correlation of SAT scores, high school grades, and socioeconomic factors. *Journal of Educational Measurement, 44*(1), 23–45.

Zwick, R., & Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement, 48*(2), 101–121.

Notes

[1] Recent research by Isaacs (forthcoming) has shown that 82.3% of comprehensive colleges and universities that offer basic writing use the results of a purchased test, such as the SAT®, Accuplacer®, or state test for placement decisions (p. 126).

[2] Elliot's point has also been articulated by Worrell, a member of the Joint Committee to revise the *Standards*: "The concept of fairness is something that anyone engaging in testing needs to think about from the beginning of the process" (F. Worell, personal communication, March 17, 2012)

[3] Measurement scholars have certainly not been remiss in engaging with legal scholarship, although discussions have often ignored shifting legal precedent (Sireci & Parker, 2006; Elul, 1998; Green, 1996), contained incorrect information (Sireci & Parker, 2006; Phillips & Camara, 2006), failed to address state and local laws (Sireci & Parker, 2006; Davis, 2006; Camilli, 2006; Pollock, 2005; Verdun, 2005; Ryan, 2003; Kidder & Rosner, 2002), or misused technical terms like disparate impact (Popham, 2012). Likewise, the legal community has been fickle in its uptake of the *Standards*, although, as Pullin (2014) pointed out, "the *Standards* have sometimes been an important influence in the outcomes of some high-visibility court cases in education and in employment" (p. 19) as well as "the more routine, ground-level decisions made in legal contexts" (p. 20).

[4] Any model that rests on the assumed stability of latent variables is suspect. The stability of latent variables overlooks not merely that students change in their knowledge, motivation, and identification with academic performance but also that their identities change over time and that those shifting identity affiliations potentially have effects on the salient latent variables (Worrell, 2014).

[5] The Center for Culturally Relevant Evaluation and Assessment at the University of Illinois has been a particularly active in the area of "culturally responsive assessment" (About CREA, n.d.). Culturally responsive assessment is a sister term to culturally responsive pedagogy, and recognizes the relevance of cultural identity in all aspects of a student's educational experience (Ladson-Billings, 1994; Nieto, 2013).

[6] Statutes and regulations are different. Statutes are bills passed by legislative bodies, such as the U.S. Congress or the New Jersey General Assembly. Regulations are detailed rules promulgated by an administrative agency, such as the U.S. Department of Education, under authority granted to the agency by a statute. Regulations outline how statutes will be interpreted and applied by an administrative agency. Both statutes and regulations have the force and effect of law.

[7] A version of this table appeared in our previous article (Poe, Elliot, Cogan, & Nurudeen, 2014). It is given here with permission in order to provide a fuller expansion of the laws than was possible in our previous publication due to space limitations.

[8] State laws differ, sometimes significantly. New Jersey's laws are used for illustrative purposes only.

[9] Since the Baxter report, much has been written about Title VII: Equal Employment Opportunity in the measurement literature (Pullin, 2013, 2014; Smith & Hambleton, 1990; Sireci & Green, 2005). Title VII makes it "unlawful to discriminate in any aspect of employment." The legal precedent for Title VII was established in the *Griggs v. Duke Power Company* case (1971) in which the Supreme Court ruled "unvalidated tests were equated with intentional discrimination" (Selmi, 2006, p. 723). In 2009 there was a twist to Title VII cases in the *Ricci v. DeStefano* case, when the city of New Haven threw out promotion test results that showed differential performance for African American candidates. White and Hispanic firefighters in New Haven challenged the city's action to throw out test results, citing disparate treatment based on race. In other words, the plaintiffs accused the city of using intentional discrimination to alleviate unintentional discrimination. The Court held that the City incorrectly discarded the test because it had not "demonstrate[d] a strong basis in evidence that, had it not taken the action, it would have been liable under the disparate-impact statute." The *Ricci* case is a good example of how test results alone, devoid of contextual factors and analysis, are insufficient to prove disparate impact.

[10] § 602 states, "each Federal department and agency which is empowered to extend Federal financial assistance to any program or activity, by way of grant, loan, or contract other than a contract of insurance or guaranty" is required to ensure that recipients are not discriminated against (1964).

[11] Differential graduation rates are often disguised in overall graduation data. Disaggregated graduation rates for students placed into remedial classes versus traditional or honors classes are rarely presented publicly but are important points of data for researchers interested in civil rights claims. For example, if graduation rates are low for students placed in basic writing (e.g., 18% in basic writing versus 40% in traditional courses), the effect of those low graduation rates are not obvious in overall graduation rates (e.g., 35%), especially if only a small number of students are required to take remedial courses versus the overall cohort. In turn, this effect is also found when data are disaggregated by race. As more students of one race are funneled into basic writing, with its lower graduation rate, the overall graduation rate for that race declines.

[12] This analysis only measures the effects of remediation in a single subject area. The cumulative effect of students placed into remediation in multiple subjects (e.g., English and Mathematics) can be even more pronounced.

[14] Self-report racial/ethnic identity can also present challenges. Likely, the most well-known challenge is the category "mixed race," which includes many Native American students. Native American self-reporting can also be challenging because self-reporting may or may not include members who are officially enrolled in an indigenous nation—for example, there are 819,000 self-identified Cherokee on the U.S. Census but only 314,000 officially enrolled Cherokee citizens.

[15] Currently, recipients qualify for reduced meals at 185% the federal poverty level and free meals at 130% the federal poverty level (Department of Education, 2015).

[16] Hispanic, of course, was not an identity designation until 1970 and even now is not considered a racial category on the U.S. Census. Instead, Hispanic origin is defined a "the heritage, nationality, lineage, or country of birth of the person or the person's parents or ancestors before arriving in the United States. People who identify as Hispanic, Latino, or Spanish may be any race" (Census, n.d.).