Routledge
Taylor & Francis Group

Educational Research and Evaluation

An International Journal on Theory and Practice

ISSN: 1380-3611 (Print) 1744-4187 (Online) Journal homepage: www.tandfonline.com/journals/nere20

Ongoing issues in test fairness

Gregory Camilli

To cite this article: Gregory Camilli (2013) Ongoing issues in test fairness, Educational Research and Evaluation, 19:2-3, 104-120, DOI: 10.1080/13803611.2013.767602

To link to this article: https://doi.org/10.1080/13803611.2013.767602

Published online: 18 Mar 2013.

Educational Research and Evaluation, 2013
Vol. 19, Nos. 2–3, 104–120, http://dx.doi.org/10.1080/13803611.2013.767602

Routledge
Taylor & Francis Group

Ongoing issues in test fairness

Gregory Camilli*

School of Education, University of Colorado at Boulder, Boulder, CO, USA

In the attempt to identify or prevent unfair tests, both quantitative analyses and logical evaluation are often used. For the most part, fairness evaluation is a pragmatic attempt at determining whether procedural or substantive due process has been accorded to either a group of test takers or an individual. In both the individual and comparative approaches to test fairness, counterfactual reasoning is useful to clarify a potential charge of unfairness: Is it plausible to believe that with an alternative assessment (test or item) or under different test conditions an individual or groups of individuals may have fared better? Beyond comparative questions, fairness can also be framed by moral and ethical choices. A number of ongoing issues are evaluated with respect to these topics including accommodations, differential item functioning (DIF), differential prediction and selection, employment testing, test validation, and classroom assessment.

Keywords: test fairness; DIF; differential prediction; employment testing

Introduction

In the attempt to identify or prevent unfair tests, quantitative analyses are often used to determine whether test items or test scores have equivalent meaning for different groups of examinees. Other instances of unfairness are not easily characterized in this manner. For example, a test may include items that are culturally insensitive, or test administration may occur in a distracting environment. This suggests a single theory or conceptual umbrella for capturing all potential threats to test fairness is not feasible. Furthermore, there are few contemporary illustrations to guide fairness evaluation in the literature showing how test bias has been prevented or rectified – other than situations in which a problematic item has been detected and removed from a test. Thus, fairness evaluation and implementation is often a pragmatic activity based on an eclectic set of analyses to determine whether due process has been accorded to either a group of test takers or an individual.

In both the individual and comparative approaches to test fairness, counterfactual reasoning is useful to clarify a potential charge of unfairness: Is it plausible to believe that with an alternative assessment (test or item) or under different test conditions an individual or groups of individuals may have fared better? An important related issue is how high the burden of proof should be for admissible counterfactual evidence. That is, should the counterfactual be probable, reasonable, or merely plausible? A third issue concerns what recourse should be offered given a successful challenge to the hypothesis that a test is fair, namely, what should the recourse be for demonstrated unfairness?

In this paper, a number of particular issues are examined including the individual and group basis for fairness evaluation, statistical methods and their limitations, employment

*Email: g.camilli@colorado.edu

testing, classroom assessment, and procedural and substantive due process. While the material on employment testing is unique to the US, these procedures may be useful for informing or examining procedures elsewhere in the world. Indeed, this is not just a controversial topic in the US, and the American experience succinctly demonstrates points of social and scientific contention.

Individual and group fairness

The 1999 *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) is intended to provide criteria for the evaluation of tests, testing practices, and the effects of test use, where the term *test* here is used to specify a broad range of assessments including particular tests, scales, inventories, and instruments. The *Standards*, for short, reflect a broad vision of test fairness:

> A full consideration of fairness would explore the many functions of testing in relation to its many goals, including the broad goal of achieving equality of opportunity in our society. It would consider the technical properties of tests, the ways test results are reported, and the factors that are validly or erroneously thought to account for patterns of test performance for groups and individuals. (p. 73)

Twelve criteria pertaining to fairness are given in the *Standards*. Nine refer to groups or subgroups, and three to individuals or aspects of individuals, suggesting the broad categories of *group* and *individual* fairness (Ferdman, 1989). Along a second dimension, six standards refer to interpretation, reporting, or use of test scores; three to differential measurement or prediction; and three to equity or sensitivity. Below, the ideas of individual and group fairness are explored in more detail. This material is key to understanding methods of fairness evaluation corresponding to the descriptions of features along the second dimension.

Individual fairness

Individual fairness requires standardized conditions of testing in which students are treated comparably. This type of fairness is denoted as equity. However, the term *comparable* is not equivalent to *equal*. For example, some students have recognized difficulties, such as dyslexia, that interfere with their test performance, while others may be English language learners. Both of these backgrounds may lead to lower performance under standard testing conditions. Thus, for their scores to be *comparable* to other test takers, accommodations may be required to offset conditions that prevent full demonstration of ability. This may consist of allowing extra time for slow readers, but may also include reading or translating a mathematics item to accommodate language issues. The goal is to prevent influences irrelevant to the test to create advantages or disadvantages that result in higher or lower test scores (Messick, 1989). In sum, if a test or test item is equitable, it is presented to individuals under impartial conditions, meaning that no individual student is favoured over another in *demonstrating what they know or understand*.

A charge of unfairness might be the argument that a different set of testing conditions would have allowed an examinee to demonstrate his or her true capacity. Of course, this counterfactual requires that the proposed altered condition (e.g., a testing accommodation such as extended time, extra breaks, or glossaries for non-native speakers of a language) would not provide an unfair *advantage*. To determine the appropriateness of an

accommodation, an experiment could be conducted in which groups of students – those who qualify for an accommodation and those who do not – are randomly assigned to accommodated and standard testing conditions. Helwig and Tindall (2003) carried out such an experiment with read-aloud accommodations for a mathematics test for both general and special education students to determine whether accommodations provided a better assessment of the proficiency of the latter. Their results suggest that accommodations can be reasonable (most of the students were diagnosed with learning disabilities), but actionable information may be ambiguous. Moreover, they a found that teachers could identify with a probability no greater than chance students who would benefit from accommodations. In a meta-analysis of 11 studies on accommodations for English language learners (ELL), Kieffer, Lesaux, Rivera, and Francis (2009) found that only one of seven typical accommodations had a relative positive impact – providing English dictionaries or glossaries. Yet, even in this case the average effect size was small (d = .15) and suggests that in some studies, accommodations increased the achievement gap between ELL and non-ELL students.

Major testing programmes nonetheless have procedures for requesting and granting accommodations. For example, the College Board requires documentation that includes a statement of the specific disability, educational history, evaluative test results (and credentials of the evaluator), explanation of a student's functional limitations, and specific accommodations requested (College Board, 2012). It can be argued along these lines that accommodations constitute a salient and pragmatic feature of test fairness beyond psychometric considerations. This is true despite the unarguable psychometric value of an accommodation in particular cases. In legal reasoning, there is a distinction between *probable cause* and *reasonable suspicion* as standards for proof. In oversimplified terms, probable cause "exist[s] where the known facts and circumstances are sufficient to warrant a man of reasonable prudence in the belief" of a conclusion (Congressional Research Service, 2006). On the other hand, reasonable suspicion of a cause–effect linkage is more than a hunch or intuition. It is a standard "considerably below preponderance of the evidence", but must flow from generalizations drawn on the basis of experience (Congressional Research Service, 2006, p. 1). With accommodations, it seems prudent and feasible to use the lesser standard: In some documented cases, it could be shown that an accommodation did enhance a student's score given an alternative test administration with no accommodation. However, this type of information is more an existence proof than a pattern with broad generalization.

Another aspect of individual fairness involves treating test takers with dignity and sensitivity. This aspect of testing may have no counterfactual stated in terms of alternative tests or test conditions: It is no defence of a charge of unfairness in this regard to argue that examinees were treated badly but equitably. It is also no defence to demonstrate equity if the objectives upon which a test is based are themselves faulty. Bouville (2008) noted that

> Giving credit to students because their name starts with a B is not any less fair than giving them credit because they have the correct answer …. Fairness of treatment is no fairness at all when it means applying rules that are themselves unfair: it merely propagates the fairness or unfairness of pre-existing rules. (p. 2.)

This is why test validation is an essential element in the analysis of fairness. The test should measure what it is intended to measure, but what the test does measure should be defensible logically and ethically. Testing conditions and test content should also avoid stereotyping, culturally offensive material, and other negative implications – regardless of whether these

factors have an effect on test performance or testing conditions are standardized. Sensitivity problems can potentially lead to statistical bias and thus faulty interpretation of test scores, but establishing a convincing causal link between insensitivity and test performance is difficult, at best (requiring an unethical experiment). Rather than attempt to adduce the effect of cultural insensitivity on test performance, it should be recognized that cultural norms may be more damaging or hurtful to individuals in other ways than underestimation of ability. The process of screening items for negative social content is called *sensitivity review*, and is a well-established feature of many testing programmes that occurs prior to operational testing.

Group fairness

It is often the case that issues of individual fairness are of concern, but statistical methodologies for test fairness often focus on group comparisons. In addition to race and ethnicity as groups of interest, other methods of classification include, but are not limited to, social class, language, urbanicity, and the like. The categorization of individuals into groups must be done with caution, and this point is illustrated below with respect to race.

In October 1997, the Office of Management and Budget (OMB) in the US released new categories for collecting data on race and ethnicity (OMB, 1997). The new racial categories established were White, Black or African American, Asian, Native Hawaiian or Other Pacific Islander, and American Indian or Alaska Native. In contrast to racial categories, several ethnic categories were designated. For designating ethnicity, the categories were Hispanic or Latino and Not Hispanic or Latino. While there was no "multiracial", individuals could select one or more races – recognizing that designations of race are becoming more complex and nuanced. Ethnicity is at least as nuanced as race, involving issues of language, race, place of origin, values, and heritage.

The charge of unfairness is often, if not inevitably linked to social bias. The argument in favour of establishing the classification in this regard is that it is the source of the social bias that creates the classification, not those who are using it to investigate potential bias. That is, a classification may indeed be socially constructed, but the central issue concerns the rationale or *causes* underlying the classification. Beyond these causes, there are also *effects* to consider, and this raises a central question in test fairness of how cause and effect are linked. Quantitative methods can improve an understanding of the link between cause and effect, but arguments solely based on the authority of statistical methods are both flawed and obfuscating (e.g., Camilli, 1993).

Statistical evidence of unfairness

To evaluate bias statistically, a counterfactual question is useful. In particular, the question is how an individual or a group would have fared on a test (or a particular test item) had the test materials or conditions not presented an obstacle to demonstrating full proficiency. The difference between the observed outcome and the counterfactual outcome constitutes the effect, but the link between cause and effect cannot be established in individual cases. Some have suggested causality cannot be determined without the possibility of random assignment to factual and counterfactual conditions (Holland, 1986). Without a controlled comparison, a scientific link below the threshold of reasonable cause, if one is to be made, must be established in a process which could be based on other criteria such as either expert judgement or legal precedent.

With classifications of individuals, an initial sign of unfairness is an average difference among different groups. However, it is important to understand the extent to which an

average test score difference is artifacts of a test or testing conditions rather than differences in true ability or proficiency. An observed difference does not necessarily imply unfairness. To give a simple example, consider the difference in average running speeds over a fixed distance for two groups (A and B), using different stopwatches. In contrast, statistical bias arises when the stopwatches do not measure time identically. In this case, at least part of the group difference in average running speed is due to the different watches. If stopwatch B runs more slowly than stopwatch A, group A is advantaged. In short, test fairness does not imply equal outcomes across individuals or groups. The purpose of a fairness investigation is to sort out whether the reasons for group differences are due to factors beyond the scope of the test (such as opportunity to learn or level of achievement) or artificially dependent on testing procedures.

There are two broad categories of analyses for examining the comparability of scores across groups: item-level analysis and prediction of a criterion measure. Both are intended to identify instances in which a test procedure is the source of group differences rather than true differences in ability. The first category is often labelled *differential item functioning* (DIF), while the second is conveniently referred to by the phrase *differential prediction*. Though the phrase test bias is often used to describe this latter category, the phrase differential prediction more accurately conveys the meaning, while the term fairness is bound to a particular use of a test in selecting candidates. It is the *selection procedure* that is described as fair or unfair, not the test itself. Differential item functioning and prediction are discussed more fully in the next sections.

Differential item functioning

Absent an external criterion, procedures were developed using the other items on the test: An item of interest had a group performance difference relatively larger than the group differences for other items. Holland and Thayer (1988) formalized the concept of relative item performance and introduced the term *differential item functioning* to convey this concept more clearly. Differential item functioning (DIF) is said to occur when examinees from groups A and B have the *same degree of proficiency* in a certain domain but different rates of success on an item. The DIF may be related to group differences in knowledge of or experience with some other topic beside the one of interest. The term DIF is synonymous with nonequivalent measurement, whereas unfairness can only be established if measurement differences are factors irrelevant to the test construct: There is no direct route from differential measurement to an inference of unfairness.

Sensitivity and specificity

Most testing programmes examine test items for group difference in measurement properties. An item is flagged when a statistically significant difference between two groups is found. However, due to statistical uncertainty, some flags are false positives. New testing techniques have resulted in a vast improvement in the reduction of false positive errors (e.g., Penfield & Camilli, 2007). However, there is tension between Type I and II errors. Type II errors occur when an item functions differentially, yet a statistical test fails to flag the item. These concerns are often described in terms of specificity and sensitivity. A test with perfect sensitivity would correctly identify all items with DIF (or no false positive): A test with perfect specificity would correctly identify all non-DIF items; that is, all items identified as having no DIF in truth have no DIF (or no false negatives). Unfortunately, there are no perfectly sensitive and specific tests for DIF. More importantly, as a

test is designed to become more sensitive, it often becomes less specific. It seems most reasonable to tip the trade-off toward sensitivity at the expense of specificity because many flagged items are not automatically rejected, but reviewed for substantive interpretations of unfairness. Only if a plausible link can be offered is an item modified or deleted. Thus, additional safeguards exist (causal linkage and minimum effect sizes) for false positives.

Explanations of DIF

Bronfenbrenner and Crouter (1983) and Bronfenbrenner (1986) argued that research that relates a macrosystem such as group identity to an outcome of interest employs a *social address model*. While such a model might correctly reveal a statistical connection between group (the "address") and individual outcomes, it would not clarify the processes that might explain the connection. In other words, the address tells you where you are, not how you got there. One significant limitation to DIF analysis results from classifying individuals by social address rather than educational histories. Implicit assumptions are made, for example, that individuals with a common race or ethnicity have the same experience when within-group variability may be greater than between-group differences (Reese, Balzano, Gallimore, & Goldenberg, 1995).

Examples of social address classifications are race, ethnicity, religion, political inclination, and so forth. The different routes to this address contain the desirable or even necessary information for explaining variance in an outcome of interest. Recall that with DIF techniques, the link between cause and effect is usually required in order to remove or modify an item from a test. A successful link entails uncovering the mediating processes which link different developmental outcomes with the address label (De Graaf, 1999). However, this information is typically not available or not used in DIF analysis. Moreover, to the degree that the classes used as the basis of a DIF analysis are heterogeneous, there is a very low chance that the mediating processes are strongly correlated with the address labels. Because many individual influences are submerged in a "social address," it is likely that the use of such classifications will not provide a very powerful means of detecting differential measurement. Consequently, the failure to reject hypotheses of differential measurement does not provide convincing support that a test item is fair. In sum, a cause–effect link can rarely be established, though modest success has been encountered with linguistically homogeneous groups (Schmitt, Holland, & Dorans, 1993). This tends to minimize the effect of Type I errors because test items are typically modified or removed only if a plausible link can be established.

Differential prediction

Differential prediction requires a criterion construct by which a person or student can be deemed successful in a given activity, and is often motivated by the need to select candidates that are more likely to be successful on the job. The idea is to select examinees using their scores on the predictor or qualifying variable. Unbiased prediction is obtained when the same prediction equation holds for two or more groups. In other words, procedure is illustrated (see Camilli & Shepard, 1994, Figure 1.1A). Despite the difference in group distributions on the predictor (disparate impact), the test has equal predictive validity for two groups because they share the same regression line. For any given qualifying score, individuals have the same expected criterion performance, regardless of group membership.

Suppose candidates are chosen for a position with a qualifying examination, and the qualifying score is used to make a prediction regarding the candidates' likelihood of

success on a criterion. For example, the Law School Admission Test (LSAT) is used to predict 1st-year grade point average (GPA) in law school. Then, as Cleary (1968, p. 115) suggested, differential prediction occurs if the criterion score predicted for a particular group is different from that of other groups given the same qualifying score (see Camilli & Shepard, 1994, Figure 1.1B). While many different models of selection based on a qualifying score have been proposed (see Camilli, 2006, for a brief review), the model most often employed is that of (a) using a common regression line across groups and (b) setting a cut-off point along the qualifying score continuum.

In the case of differential prediction, a common regression line used to select candidates for college or a job would lead to overprediction of performance for one group and underprediction for another, leading to an *unfair selection process*. An initially obvious solution to this problem is to use different prediction equations. However, such a solution is not feasible: Using different qualifying scores for different groups would itself be perceived as unfair outside of a technical community. More important, as a number of authors have argued (Borsboom, Romeijn, & Wicherts, 2008; Cronbach, Rogosa, Floden, & Price, 1977; Lord, 1960; Millsap, 1998, 2007), test scores with measurement error – which means all tests – will have different prediction equations for two groups with mean separation, even when there is no DIF in test items. Camilli, Briggs, Sloane, and Chiu (2013) further showed that differential selection can be quite large in some cases with a test composed of items with equivalent measurement properties.

A related issue with criterion prediction is whether a test may underrepresent the qualities desired for success (construct deficiency). However, this is more of an assumption underlying an analysis of differential prediction (along with construct contamination) than an issue to be investigated with differential prediction. Content validity along with expert review and job analysis are often the methods for examining the latter two qualities of a test.

Other fairness perspectives

Above, a number of limitations were examined for both differential item functioning and differential prediction. A broad perspective of test fairness requires coverage of still more issues in which test use is consequential to individuals. One such issue that stirs controversy is selection in employment testing. A perspective on the current situation in the United States is given, but the essential elements of this debate are likely to be relevant internationally. Second, test fairness is examined in the framework of test development. Third, a perspective on classroom testing is given, recognizing that this is the most prevalent kind of testing in the US and elsewhere.

Legal issues in employment testing

For some, fairness requires treating people as individuals, and for others, fairness requires taking into account the collective representations that matter in society. Levin (2003) frames this fairness debate in terms of a distinction between the individual and group perspective:

> Proponents of the individualistic perspective argue that it is unfair to pay attention to ethnicity because ethnic group memberships should not influence the opportunities and outcomes of individuals in society. Proponents of the group perspective, on the other hand, argue that it is unfair not to take ethnicity into account because of the power differentials that exist between ethnic groups in society. According to this latter perspective, ignoring ethnic group

> membership obscures the significant ways in which these power differentials influence the opportunities and outcomes of members of different ethnic groups. (p. 8)

Both aspects of fairness must be considered in an evaluation of employment selection. The message from social address theory should not be taken to imply that the "social labels" have no inherent meaning; indeed, stereotyping and historical discrimination are directed precisely toward the "label" rather than the individual. In turn, this provides the starting point for a legal due-process perspective.

Title VII of the Civil Rights Act of 1964 concerns discrimination in employment practices, and the Equal Employment Opportunity Commission (EEOC) was created to provide guidance and enforcement regarding Title VII. Section 703(a)(2) declares it unlawful to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin. The EEOC in 1966 interpreted Title VII discrimination to consist of employment practices intended to discriminate or to treat people of protected status differently from others, but also includes practices having a harmful affect on members of certain groups (Hartigan & Wigdor, 1989).

Title VII provides for challenges to employment practices that adversely affect such groups – whether the outcome was intentional or not (Griggs v. Duke Power Company, 1971). Adverse impact, which is a measurable outcome derived from the Title VII term "adversely affect", has come to denote the selected proportions of people from groups with protected status relative to unprotected or majority groups. Adverse or disparate impact in employment decisions has been defined as a substantially different rate of selection that creates an imbalanced workforce with respect to a group with protected status. Note that the adjective *disparate* rather than *adverse* impact is used in the Civil Rights Act of 1991, PL 102-166). In particular, the usual threshold for adverse impact is established if the selection rate for one group is less than 80% (known as the 4/5 rule) of that for the group with the highest selection rate according to the *Uniform Guidelines on Employee Selection Procedures* (EEOC, 1978). However, courts have not rigidly interpreted statistical criteria for demonstrating disparate impact, and disparate impact by itself is not a sufficient basis for establishing violation of equal protection. Thus, the 4/5 rule is a guiding principle, and courts have also admitted evidence based on both statistical and practical significance (e.g., Gregory v. Litton Industries, 1972).

The demonstration of disparate impact establishes the grounds for legal challenge, i.e., a prima facie argument, and the effect of Title VII was to place the burden of proof on employers in the context of adverse impact given successful demonstration of disparate impact (e.g., demonstrating the 4/5 rule). To defend practices that result in alleged disparate impact, defendants may challenge the plaintiff's statistics or present three types of information on test validation:

(1) Criterion-related validity: A statistical correlation between scores on a selection test and job performance on a sample of workers.
(2) Content validity: A demonstration that the content of a selection procedure is representative same of important job-performance tasks.
(3) Construct validity: A demonstration that (a) a selection procedure measures a construct, and (b) the construct represents an important capacity for successful job performance.

It is the test user's responsibility to determine how to satisfy these three criteria, and this may include the use of previous validity studies (though the procedure for selecting previous studies must be consistent with the validation strategy). The third strategy above of construct validity is not well-developed in the content of the *Uniform Guidelines*. Construct evidence might take the form of showing that in previous validity studies test scores (for the test in question) correlated with the same requisite work behaviours in the context of different jobs, whereas concurrent validity evidence should concern correlations in the context of jobs with substantially the same work behaviours.

A selection practice should be justified by demonstrating it is job related and consistent with business necessity (Civil Rights Act of 1964), though there is some uncertainty of whether the word *consistent* is interchangeable with *necessary* (Grover, 1996). Arguments based on the "rational" relationship between a selection procedure and job performance are an insufficient defence (Washington v. Davis, 1976). If another feasible selection procedure is available that results in less disparate impact, that alternative procedure should be included in the validity study. If other alternatives result in even less disparate impact, the burden of evidence is on the plaintiff to uncover and demonstrate these practices. In any event, the Civil Rights Act imposes liability if the defendant refuses to adopt a validated selection practice that results in less disparate impact, if it can be identified.

Underneath the surface of this terse description, controversy surrounding the application of current employment selection policies is rampant. This is revealed by an exchange of views on a paper by McDaniel, Kepes, and Banks (2011b), "The Uniform Guidelines Are a Detriment to the Field of Personnel Selection", which appeared in the journal *Industrial and Organizational Psychology* (Volume 4, 2011).

McDaniel et al. (2011b) argued that the *Uniform Guidelines* encourage the "use of selection practices unsupported by scientific evidence" (p. 495). They further maintained that disparate impact is an unavoidable fact and that differential prediction does not exist. Perhaps most important, they also argue that with the "scientific" advance established by validity generalization, there is little need, if any, for local validation studies (that is, situational specificity is largely a myth) in response to Title VII challenges: "From the perspective of scientific knowledge, meta-analytic evidence largely eliminates the need for local validation studies" (p. 500). The bottom line in this argument appears to be the claim that disparities in selection ratios (e.g., the 4/5 rule) "should not generally trigger federal intervention in personnel selection practices" (p. 495). Consistent with this position, Barrett, Miguel, and Doverspike (2011) advocated removing the disparate impact theory of discrimination entirely from Title VII policies.

McDaniel et al. (2011b) identified three "tactics" used against the modification of the *Uniform Guidelines*:

> First, employers can document the costs associated with complying with the Uniform Guidelines. These include labor and other monetary costs associated with defending employee selection systems. There are also economic costs associated with using lower validity selection measures in hopes of reducing adverse impact …. Second, employees of federal regulatory agencies, human resources consultants, and labor lawyers seeking to preserve their jobs can manufacture uncertainty about scientific findings. If the price is right, one can find a "scientist" to testify to almost anything. Third, regulatory agencies and other interested parties (e.g., consultants, lawyers, and expert witnesses) can engage in delay tactics (e.g., litigation, requiring parallel studies and fighting over access to raw data) to avoid revising the Uniform Guidelines. (p. 509)

McDaniel et al. (2011b, p. 507) expressed the hope that given President Obama's mixed-racial heritage, an Obama-endorsed congressional effort to force a revision of the

Uniform Guidelines is less likely to be labelled as racially motivated. Finally, they faulted the *Uniform Guidelines* for not addressing the diversity-validity dilemma (pp. 506–507).

One would be incorrect to conclude that this point of view represented an endpoint along a continuum of disagreement. In more vigorous response, Sharf (2011) labelled the *Uniform Guidelines* as a "tool of political advocacy" and "the civil rights bar's hostility to job-related selection procedures that adversely affect preferred groups is based on their stealthy redistributive advocacy of equal employment" (p. 537). Sackett (2011) viewed the current *Uniform Guidelines* as a policy/political document rather than a scientific/professional document and maintains "it is thus inappropriate to rely on them as the basis for one's professional opinion" (p. 545), and further, the

> appropriate bases for testimony in selection cases are the peer-reviewed literature and the consensus documents that reflect concepts and procedures that have gained acceptance in the scientific community, namely, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) and the *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003). (p. 545)

Sackett also clarified that the key rationale for testimony inconsistent with the Standards or Principles is to provide a recommendation that is "is not a scientific claim, but rather some other factor, such as avoiding scrutiny from a regulatory agency" (p. 546).

In contrast to the views above, Tonowski (2011) argued that "One purpose of the *Uniform Guidelines* is clearly to address public policy regarding the use of tests and other selection procedures" (p. 527). Indeed, McDaniel et al. (2011a) suggested this should be the case in identifying the limitations of the *Uniform Guidelines* in addressing the diversity-validity dilemma. Tonowski also pointed out that a set of revised *Standards* will shortly be published, and the *Principles* may have to be revised accordingly prior to an attempt to revise the *Uniform Guidelines*. In addition, it seems likely that an agency or committee charged with revision tasks would itself require a complex selection process for members, and it is not clear whether such a committee could arrive at a consensus on key issues. However, Reynolds and Knapp (2011) suggested a broad outline for projects that might usefully serve to guide revision.

Outzz (2011) argued "there is credible evidence refuting that situation specificity is a "myth" (p. 529). He noted that "Ironically, the use of meta-analytic methods has revealed that other assessment tools or combinations of tools can result in validity equal to, if not higher than that for cognitive ability tests and produce less adverse impact…." (p. 531). In the *Principles* (Society for Industrial and Organizational Psychology [SIOP], 2003) in contrast, it is stated that generalized evidence can be used to support test validation claims given a "compelling argument for its application to the situation of interest" (Outzz, 2011, p. 530). In fact, it is explicitly stated in the *Standards* (AERA, APA, & NCME, 1999) that "The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used" (p. 11) (also see *Standards* 1.1 and 1.20).

Brink and Crenshaw (2011) provided a counter-pole to the views presented in the focal paper. They argued that many of the criticisms lodged by McDaniel et al. (2011b) are adequately addressed in the *Uniform Guidelines*, but also in questions and answers (Q&As) provided by the Equal Employment Opportunity Commission (EEOC, 1979, 1980), which McDaniel et al. omitted from their discussion. The thesis of Brink and Crenshaw is that

> the majority of their arguments [McDaniel et al., 2011a] are either incorrect assertions, beliefs presented without basis in established fact, or trivial in terms of implications for practitioners –

> none are significant enough for the *Uniform Guidelines* to be considered a detriment or rise to the level of calling for the *Uniform Guidelines* to be rescinded. (p. 547)

A selection of the critiques offered by Brink and Crenshaw includes the role of validity generalization (VG), the gap in selection test means for White and minority job candidates, and the roles of science and law in employment regulations. First, they argued that both the constructs and methods of measurement must be similar for VG to apply to a local test, and "Though the Principles address VG, they do not support its cavalier use" (p. 548). Second, they noted that while McDaniel et al. (2011a) provided evidence of group differences on many important selection variables, one was notably absent from their list: job performance. This is important to consider because the Black versus White gap in performance is considerably smaller than many other predictors of job success. Exclusive use of cognitive measures may focus too narrowly on the knowledge and skills required for effective job performance. Finally, Brink and Crenshaw (2011) argued that "science/practice and law should mutually influence one another and that a perfect match between science and law is a noble goal but difficult to attain the idea", and "it is egocentric to suggest that science takes precedence over law" in federal employment regulations (p. 552).

Despite the stark disagreement over the role and use of the *Uniform Guidelines* among industrial and organizational psychologists, there is agreement on some issues. First, practical significance (effect size) and statistical significance have been conflated in some legal proceeding (Jacobs, Deckert, & Silva, 2010). Some clarification in employment regulations would be helpful despite the fact that a cutoff for establishing a disparate impact review is a social construction rather than a scientific standard. Second, the method of *synthetic validity* provides a promising tool because it has the potential to combine generalized evidence and local context into a principled validity argument (Mead & Morris, 2011). However, as Russell (2010) noted, citing Milkovich and Newman (2007), it is estimated that "only about 20% of a typical firm's jobs have external labor markets" (p. 340). This poses a strong challenge for selection research on external evidence. Third, it would be useful to clarify whether the enduring observed score gaps on cognitive variables is an appropriate legal basis for rejecting the disparate impact theory of discrimination (Hanges, Aiken, & Salmon, 2011). After all, this argument would itself hinge to some degree on a legally suspect classification.

It is unlikely that any resolution on these matters is forthcoming, mainly because it is not clear what alternatives to current employment regulations might be palatable. Barrett et al. (2011) proposed the vague recommendation that "researchers should collaborate and engage in a professional discussion of ethnic group differences in test scores" (p. 535). And as McDaniel et al. (2011a) observed, the courts are unlikely to change their practices without fundamental changes to the Civil Rights Act of 1991. Yet, these authors other than seeking to "foster constructive debate" (2011a, p. 570), admit "no substantive experience in how to resolve the unfortunate situation with the *Uniform Guidelines*" (2011b, p. 509). Ultimately, the manner in which McDaniel et al. (2011a) responded, in their rejoinder, to the challenge of Brink and Crenshaw (2011) suggests little common ground: "the commentary [Brink and Crenshaw] appears ideologically driven, in part, because it assumes that it is acceptable for governmental officials to manipulate and deny research findings to advance political goals" (p. 569), and "We encourage individuals and professional organizations to continue this debate until detriments to the professional practice of personnel selection have been neutralized" (p. 570).

It is important to recognize that the applicability of Title VII challenges was reduced in 2001 when the Supreme Court ruled in Alexander v. Sandoval (2001) that disparate impact

arguments could not be brought to federal courts by individual citizens (right of private action), as had been the case for the previous 35 years. This has effectively rendered disparate impact viable only in the context of Office of Civil Rights administrative enforcement actions (Welner, 2001), and this opinion may reduce the number of venues in which evidence, psychometric or otherwise, of disparate impact is salient. However, as dire as this situation sounds, a discrimination charge in employment selection is most likely to be unsuccessful. For example, in 2010 about 73,000 private-sector Title VII discrimination charges were filed for which the EEOC found no reasonable cause in about 70%; moreover, about 18% were closed for administrative reasons. Reasonable cause for the belief that discrimination had occurred was found by the EEOC in only about 5% of the cases, though charging parties had meritorious resolutions in about 20% of the cases leading to 229.8 million USD in benefits. Thus, discrimination charges do not have resolutions in which the charging parties benefit (meritorious resolutions) about 9 times out of 10. On the other hand, meritorious resolutions resulted in an average of about 16 thousand USD per charge in benefits (EEOC, 2010).

Test development

Though test validation was briefly discussed above in the legal perspective, more details on the test development process are useful. Collecting validity evidence is a cornerstone of test development which can be broken down into several distinct components including construct, content, concurrent, and predictive. However, the most common, if not convincing, evidence of validity for large-scale assessment is built into procedures for developing a test. Many problems associated with test validity may be avoided in the pre-operational stages.

Test items are written according to an overall plan in large-scale assessment programmes. This plan is sometimes called a content-by-process matrix, or more simple a test blueprint. Often, this plan is guided by a set of standards based on a formal curriculum; this implies that a good test is tightly aligned to both a curriculum and a test blueprint. The latter are also aligned in a coherent system of educational practices. Though standards documents typically do not provide enough detail for writing specific test items, intermediate tools can be used to articulate test blueprints with standards documents. For each "standard", an organization may develop a set of topics and indicators within those topics describing particular skills or content. In other words, indicators and other tools, as policy interpretations of a curriculum, are used to provide explicit information to item writers. For example, several hundred indicators may be defined for an assessment, where each indicator is linked to a standard within a content area for a test composed of 40–50 items. In the alignment process, content expertise and expert judgement are essential tools, and test validation is partially accomplished by setting forth both the design decisions and expert review process in a formal document often referred to as a technical manual or report.

After development, test items go through several stages of screening. Initially, items are reviewed by substantive experts to identify intrinsic ambiguities. They are further reviewed by expert panels for sensitivity to social and cultural content. In a second stage, items are piloted of samples of students for a number of reasons. An example of a problem that might arise with respect to fairness is the use of the word "snow" with student population having very little experience with snow. This problem might be detected with either sensitivity review or possibly DIF analysis. Test items often undergo a third stage of pretesting to determine if groups of items work together as expected. In short, a number of steps take place that in part address fairness issues that include match to the test blueprint and content standards, suitable cognitive processes, unfamiliar or insensitive language,

inappropriate difficulty, and so on. Thus, fairness rather than being a property of a test is more accurately viewed in terms of the coherence of the claims within an educational system. Other steps to enhance validity occur in test administration, scoring, and score reporting.

Not all problems can be resolved prior to formal administration. As summarized by Kane (2010),

> A basic case for the validity of the proposed interpretation and use (i.e., the generic interpretive argument) is typically made during test development. This development stage tends to have a confirmationist bias, because it is an integral part of developing an assessment that is designed to support certain interpretations and uses…. (p. 181)

Nonetheless, there are many steps in the development process designed to identify and prevent problems related to fairness issues.

Classroom assessment

Shepard (2006) defines formative assessment as "assessment carried out during the instructional process for the purpose of improving teaching or learning", while summative assessment "refers to the assessments carried out at the end of an instructional unit or course of study for the purpose of giving grades or otherwise certifying student proficiency" (p. 627). Any assessment that eventually affects a grade can be viewed as summative, and therefore many assessments, including standardized tests, can share both summative and formative purposes (Brookhart, 2003, 2004).

Classroom assessment is by far the most prevalent type of testing, yet most formal methods of test fairness are restricted to large-scale assessment. As a rule, classroom tests have a short life cycle, and it would not be feasible to techniques of differential measurement or prediction. Camilli (2006) suggested that a number of categories could be considered in evaluating fairness in classroom testing including clear and reasonable assessment criteria, the strength of the link between assessment and instruction, opportunity to learn, sensitivity of assessment procedures to cultural and religious differences, and the use of multiple measures. Moreover, because the instructor is the grading authority in the classroom, grading can be a type of modelling of both learning and self-assessment of students. When fair assessment procedures are internalized, they become a model process for the students' participation in their communities and a larger democratic culture.

Discussion

A number of current topics related to test fairness have been briefly sketched in this paper. First, a number of factors prevent DIF methods from being definitive including inexact group classification, cause–effect linkages, and the issues of sensitivity and specificity. The expectations that DIF analysis can lead to fairer tests should remain appropriately modest: It is but one step in a quality control procedure. The argument that DIF analysis alone is adequate for creating a fair test is not accepted within the psychometric community.

Second, significant limitations also exist with respect to differential selection. Even with a test composed of items that have equivalent measurement properties in two groups, differential prediction occurs to some degree due to measurement error when two groups differ in average score on a qualification test. It is well known that in the presence of group separation on the predictor variable, a common prediction line typically gives, on average, a slight selection advantage to members of lower scoring groups.

Third, according to the legal perspective, an extant selection test sets a baseline for evaluating whether other tests exist that have less disparate impact. Statistical analysis may have a role to play in the comparison of two alternative tests, but if the tests contain different amounts of measurement error, such an analysis cannot be definitive. Likewise, analyses are limited by the suitability of the criteria to be predicted. The choice of selection criteria is instead determined in a proceeding in which various types of validity information are offered and challenged. It is at this stage that an expert investigation of test content may be undertaken to determine whether a test is appropriately targeted to job-necessary behaviours, and whether the test accurately represents the indicators relevant to such behaviours.

Should there be frameworks for considering fairness issues such as Toulmin's argument structure (2003), or test validation? The approach taken in this paper is informed by the potential outcomes model (Holland, 1986) in which a counterfactual question is proposed: Would the assessment outcome have been different under other circumstances – including different tests, accommodations, or other altered conditions of testing? It is a question easily asked, but answered with much difficulty, if at all. Empirical investigation may provide clues for this purpose, if not useful evidence, but the fairness argument often takes place on procedural grounds due to unavoidable constraints in establishing cause–effect relationships and identification of a consensual criterion. According to Kane (2010)

> *procedural fairness* can be said to require that all test takers be treated in essentially the same way, that they take the same test or equivalent tests, under the same conditions or equivalent conditions, and that their performances be evaluated using the same (or essentially the same) rules and procedures. (p. 178)

Procedural fairness, especially as procedural due process, should also be extended to how tests are developed and how testing is aligned to other features within an educational system. This is especially evident in the case of classroom assessment.

On the other hand, Kane (2010) describes *substantive due process* as requiring that procedures to be applied are "reasonable in general and in the context in which they are applied" (p. 178). He extends this notion to *substantive fairness* as requiring that the "score interpretation and any test-based decision rule be reasonable and appropriate, and in particular, that they be equally appropriate for all test takers (at least roughly)" (p. 178–179). There is no doubt that the purpose of a cognitive test should be to assess some valuable aspect of cognitive behaviour, but as a whole this definition is a narrower framing of substantive due process than the counterfactual question "Is it reasonable to believe that with an alternative procedure an individual or group of individuals may have fared better?" The word *reasonable* can refer to a conclusion drawn by a "prudent and cautious" person. While the lower standard of reasonable suspicion is often applied in test development and accommodations, a choice between reasonable cause and suspicion is required. But these standards are considerably lower than that of scientific evidence, which is difficult to establish outside the realm of randomized experiments. In any case, there should be a structure, if not a theory, for a fairness argument, and it should be recognized that warrants for linking evidence to potential causes are not exclusively scientific.

References

Alexander v. Sandoval, 121 S. Ct. 1511 (2001).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

Barrett, G. V., Miguel, R. F., & Doverspike, D. (2011). The uniform guidelines: Better the devil you know. *Industrial And Organizational Psychology*, *4*, 534–536.
Borsboom, D., Romeijn, J. W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, *13*, 75–98.
Bouville, M. (2008). *The obsession with exam fairness*. Retrieved from http://www.mathieu.bouville.name/education-ethics/Bouville-exam-fairness.pdf
Brink, K. E., & Crenshaw, J. L. (2011). The affronting of the uniform guidelines: From propaganda to discourse. *Industrial and Organizational Psychology*, *4*, 547–533.
Bronfenbrenner, U. (1986). Ecology of the family as a context for human development: Research perspectives. *Developmental Psychology*, *22*, 723–742.
Bronfenbrenner, U., & Crouter, A. C. (1983). The evolution of environmental models in developmental research. In P. H. Mussen (Series Ed.) & W. Kessen (Vol. Ed.), *Handbook of child psychology: Vol. 1. History, theories, and methods* (4th ed., pp. 357–413). New York, NY: Wiley.
Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, *22*, 5–12.
Brookhart, S. M. (2004). *Grading.* Upper Saddle River, NJ: Pearson Education.
Camilli, G. (1993). The case against item bias techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 397–417). Hillsdale, NJ: Lawrence Erlbaum Associates.
Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport, CT: American Council on Education/Praeger.
Camilli, G., Briggs, D. C., Sloane, F. C., & Chiu, T.-W. (2013). Psychometric perspectives on test fairness: Shrinkage estimation. In K.F. Geisinger (Ed.-in-Chief), B. A. Bracken, J. F. Carlson, J. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Assoc. Eds.), *APA handbooks in psychology: APA handbook of testing and assessment in psychology: Volume 3. Testing and assessment in school psychology and education*. Washington, DC: American Psychological Association.
Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. Hollywood, CA: Sage.
Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(k)(1)(A)(i). (1964).
Civil Rights Act of 1991, S. 611, 102nd Cong. (1991).
Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124.
College Board. (2012). *Accommodations*. Retrieved from http://professionals.collegeboard.com/testing/ssd/accommodations
Congressional Research Service. (2006, January 30). *Probable cause, reasonable suspicion, and reasonableness standards in the context of the fourth amendment and the foreign intelligence surveillance act* (Memorandum to the Senate Select Committee on Intelligence). Retrieved from http://www.fas.org/sgp/crs/intel/m013006.pdf
Cronbach, L. J., Rogosa, D. R., Floden, R. E., & Price, G. G. (1977). *Analysis of covariance in nonrandomized experiments: Parameters affecting bias* (Occasional Paper). Stanford, CA: Stanford Evaluation Consortium, Stanford University.
De Graaf, J. W. (1999). *Relating new to old: A classic controversy in developmental psychology* (Doctoral dissertation). Groningen, The Netherlands: Regenboog.
Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register*, *43*, 38290–39315.
Equal Employment Opportunity Commission. (1979). Adoption of questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. *Federal Register*, *44*, 11996.
Equal Employment Opportunity Commission. (1980). Adoption of additional questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. *Federal Register*, *45*, 29530.
Equal Employment Opportunity Commission. (2010). *Title VII of the Civil Rights Act of 1964 Charges*. Retrieved from http://www.eeoc.gov/eeoc/statistics/enforcement/titlevii.cfm
Ferdman, B. M. (1989). Affirmative action and the challenge of the color-blind perspective. In F. A. Blanchard & F. J. Crosby (Eds.), *Affirmative action in perspective* (pp. 169–176). New York, NY: Springer-Verlag.
Gregory v. Litton Industries, 472 F.2d 631 (9th Cir., 1972).

Griggs v. Duke Power Company, 401 U.S. 424 (1971).
Grover, S. (1996). *The business necessity defense in disparate impact discrimination cases* (Faculty Publications, Paper 19). Retrieved from http://scholarship.law.wm.edu/facpubs/19
Hanges, P. J., Aiken, J. R., & Salmon, E. D. (2011). The devil is in the details (and the context): A call for care in discussing the uniform guidelines. *Industrial and Organizational Psychology*, *4*, 562–565.
Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues and the General Aptitude Test Battery.* Washington, DC: National Academy Press.
Helwig, R., & Tindal, G. (2003). An experimental analysis of accommodation decisions on large-scale mathematics tests. *Exceptional Children*, *69*, 211–225.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.
Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.
Jacobs, R., Deckert, P. J., & Silva, J. (2011). Adverse impact is far more complicated than the uniform guidelines indicate. *Industrial and Organizational Psychology*, *4*, 558–561.
Kane, M. (2010). Validity and fairness. *Language Testing*, *27*, 177–182.
Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D.J. (2009). Effectiveness of accommodations for English Language Learners taking large-scale assessments. *Review of Educational Research*, *79*, 1168–1201.
Levin, S. (2003). Social psychological evidence on race and racism. In M. Chang, D. Witt, K. Haikuta, & J. Jones (Eds.), *Compelling interest: Examining the evidence on racial dynamics in higher education in colleges and universities* (pp. 97–125). Stanford, CA: Stanford University Press.
Lord, F. M. (1960). Large-scale covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, *55*, 307–321.
McDaniel, M. A., Kepes, S., & Banks, G. C. (2011a). Encouraging debate on the uniform guidelines and the disparate impact theory of discrimination. *Industrial and Organizational Psychology*, *4*, 566–570.
McDaniel, M. A., Kepes, S., & Banks, G. C. (2011b). The *Uniform Guidelines* are a detriment to the field of personnel selection. *Industrial and Organizational Psychology*, *4*, 494–514.
Mead, A. D., & Morris, S. B. (2011). About babies and bathwater: Retaining core principles of the uniform guidelines. *Industrial and Organizational Psychology*, *4*, 554–557.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.
Milkovich, G., & Newman, J. (2007). *Compensation* (9th ed.). New York, NY: McGraw-Hill.
Millsap, R. E. (1998). Group difference in intercepts: Implications for factorial invariance. *Multivariate Behavioral Research*, *33*, 403–424.
Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*, 461–473.
Office of Management and Budget. (1997, October 30). *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity* (Federal Register Notice 62FR58782-89). Washington, DC: Author.
Outtz, J. L. (2011). Abolishing the uniform guidelines: Be careful what you wish for. *Industrial and Organizational Psychology*, *4*, 526–533.
Penfield, R., & Camilli, G. (2007). Test fairness and differential item functioning. In C. R. Rao (Ed.), *Handbook of statistics: Volume 26. Psychometrics* (pp. 125–167). Amsterdam, The Netherlands: Elsevier.
Reese, L., Balzano, S., Gallimore, R., & Goldenberg, C. (1995). The concept of educación: Latino family values and American schooling. *International Journal of Educational Research*, *23*, 57–81.
Reynolds, D. H, & Knapp, D. J. (2011). SIOP as advocate: Developing a platform for action. *Industrial and Organizational Psychology*, *4*, 540–544.
Russell, C. J. (2010). Better at what? *Industrial and Organizational Psychology*, *3*, 340–343.
Sackett, P. R. (2011). The uniform guidelines is not a scientific document: Implications for expert testimony. *Industrial and Organizational Psychology*, *4*, 545–546.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sharf, J. C. (2011). Equal employment versus equal opportunity: A naked political agenda covered by a scientific fig leaf. *Industrial and Organizational Psychology*, *4*, 537–539.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (pp. 623–646). Westport, CT: Praeger.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Tonowski, R. F. (2011). The uniform guidelines and personnel selection: Identify and fix the right problem. *Industrial and Organizational Psychology*, *4*, 521–525.

Toulmin, S. (2003). *The uses of argument* (2nd ed.). Cambridge, UK: Cambridge University Press.

Welner, K. (2001). Alexander v. Sandoval: A setback for civil rights. *Educational Policy Analysis Archives*, *9*(24). Retrieved from http://epaa.asu.edu/epaa/v9n24.html

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