Educational Measurement
ISSUES AND PRACTICE
NCME national council on measurement in education

Educational Measurement: Issues and Practice
Winter 2019, Vol. 38, No. 4, pp. 34–41

Fairness in Measurement and Selection: Statistical, Philosophical, and Public Perspectives

Rebecca Zwick, *Educational Testing Service*

Selection decisions have a major impact on our education, occupation, and quality of life, and the role of standardized tests in selection has always been a source of controversy. Here, I consider various definitions of fairness in measurement and selection—those emerging from within educational measurement and statistics, those from philosophy, and finally, those from the public. I use examples of public challenges to selection practices to illustrate the fact that technical and philosophical definitions of fairness do not align well with public concerns. I emphasize the importance of promoting awareness of existing standards, advocating for the fair use of testing and selection practices, and communicating in a candid and straightforward way when engaging with test takers and test users.

Keywords: admissions, fairness, selection

Introduction

Selection decisions have a major impact on the education we receive and on our subsequent occupations and quality of life. It is not surprising, then, that the role of standardized tests in selection has always been a source of controversy, both in academic circles and among the general public. Here, I consider various definitions of fairness in measurement and selection—those arising from within the field of educational measurement and statistics, those from philosophy, and finally, those from public discourse. I use examples of public challenges to selection practices to illustrate the fact that technical and philosophical definitions of fairness, while intellectually interesting and potentially valuable as guideposts for data-analytic investigations of fairness, do not align well with public concerns and are not typically useful to measurement professionals who strive to improve communications with test users, test takers, and policymakers. I conclude with some proposals for improving communication with the public about measurement. In particular, I offer recommendations for responding to criticisms and challenges to the use of tests in selection.

Definitions from the Fields of Educational Measurement and Statistics

Can psychometrics or statistics offer a canonical definition of fairness in measurement and selection? Concerns about fairness feature prominently in the educational measurement literature. Fairness is addressed in one of the three foundational sections of the most recent edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), along with validity and reliability, and was the subject of an extensive chapter (Camilli, 2006) in the most recent edition of *Educational Measurement* (Brennan, 2006), described as "the bible in its field." However, no universally accepted definition of test fairness has ever emerged. As acknowledged in the *Standards*, "it is possible that individuals endorse fairness in testing as a desirable social goal, yet reach quite different conclusions about the fairness of a given testing program" (AERA, APA, & NCME, 2014, p. 49).

The closest the field has come to a universal definition of fairness in selection is the principle suggested by Humphreys (1952) and detailed and tested by Cleary (1968). Her definition of test bias was initially endorsed not only by educational researchers, but by industrial psychologists and authors of prominent textbooks on psychological testing (Schmidt & Hunter, 1974).

> A test is biased for members of a subgroup ... if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup ... The test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair," particularly if the use of the test produces a prediction that is too low. If the test is used for selection, members of a subgroup may be rejected when they were capable of adequate performance (Cleary, 1968, p. 115).

In her study, Cleary examined the use of the SAT to predict the college grades of Black and White students at three integrated colleges. In particular, she sought to determine whether using the SAT as a predictor led to unfairly low predictions of college grades for Black students. She indicated that "because high school rank-in-class is generally used with the SAT for the prediction of grades, rank-in-class was included in the analysis when possible" (p. 116). At two of the schools, regression slopes and intercepts for Black students

Rebecca Zwick, Educational Testing Service; rzwick@ets.org. This paper is based on my NCME presidential address, delivered on April 7, 2019 in Toronto, Canada. I appreciate the comments received from Randy Bennett, James Carlson, and Michael Kane.

did not differ to a statistically significant degree from those for White students. In the third school, intercepts differed, but not in the expected direction. Using the regression equation based on White students or the equation based on the White and Black samples combined led to overprediction of the college grades of Black students: The predicted average college GPA was higher than the actual average GPA.[1]

In the following years, several additional studies that used Cleary's regression approach showed that the college grades of Black students were overpredicted when SAT scores and high school grades were used as predictors, and an analogous phenomenon was observed in the personnel selection arena (Flaugher, 1970; Linn, 1973). Perhaps because of the unexpectedness of this finding, Cleary's definition was subjected to further scrutiny, which continues today.

In general, critiques of the Cleary model point out that regressions can differ across groups for many reasons; evidence of differential prediction does not imply test bias. For example, the criterion variable (typically, a grade, in the case of analyses of admissions) could itself be biased (Linn, 1973). In addition, as noted by Linn (1984), regression models that are identical in the population can be rendered unequal via selection, and different regressions can be made equal. This observation is important because studies of differential prediction are typically conducted after selection has already taken place.

Einhorn and Bass (1971) pointed out that two groups can have identical regression equations but unequal residual variances. In this situation, members of two groups will have the same predicted criterion score for all values of the predictor, but will not have the same probability of exceeding a cut point on the criterion.[2]

Also, measurement error in the predictors can produce a conclusion of unequal regressions. In fact, the overprediction of the criterion variable for lower-scoring groups, a finding that has been replicated consistently over the last 50 years, can be fully explained by measurement error (Linn & Werts, 1971), though other phenomena may be involved as well (Zwick, 2017). Some researchers have argued that studies of differential prediction should routinely incorporate adjustments for measurement error (e.g., Culpepper, Aguinis, Kern, & Millsap, 2019), but it is not obvious that this is the case. Selection decisions are made on the basis of the observed predictors, not the underlying constructs. When the goal is to study the practical effects of using a test for selection, it seems appropriate to focus on the properties of the observed scores on that test.

Yet another issue to consider when studying differential prediction is model underspecification. The regression models that are considered in most admissions studies are very sparse, typically including only admissions test scores and previous grades as predictors because these are the factors most often used by the institutions that process student applications. However, we know that there exist other predictors of subsequent academic performance, and that the distributions of these predictors may differ across key demographic groups. Under these circumstances, systematic prediction errors will be reduced if these supplementary predictors are included in the regression model (Linn, 1973; Linn & Werts, 1971).

For example, in a study of about 71,000 students who were undergraduates in 2006, Igor Himelfarb and I found that college first-year grade-point average (GPA) was overpredicted for Black and Latino students when only SAT scores and high school grades were used as predictors (especially if high school grades were used alone), but that overprediction was reduced when a measure of high school–level socioeconomic status was included in the model (Zwick & Himelfarb, 2011). As another example, consider the underprediction of women's first year college GPAs—another typical research finding (Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008; Young, 2004; Zwick 2002). In some studies, underprediction was reduced by including additional predictors in the model, such as the student's college major (Pennock-Román, 1994), an index of the grading severity of the student's college courses (Ramist, Lewis, & McCamley-Jenkins, 1994), or measures of the student's academic preparation and studiousness (Stricker, Rock, & Burton 1993). These results raise the question of what predictors should be included when using a regression approach for judging the fairness of a selection process.

It is telling that the 2014 edition of the *Standards for Educational and Psychological Testing* departs from Cleary's perspective on fairness in which it does not identify differential prediction as evidence of test bias. Instead, the *Standards* says that

> [t]he term *predictive bias* may be used when evidence is found that differences exist in the patterns of associations between test scores and other variables for different groups, bringing with it concerns about bias in the inferences drawn from the use of test scores. (AERA, APA, & NCME, 2014, p. 51)

This is a tacit acknowledgment that studies of differential prediction, while they may be useful, fall short of providing conclusive evidence about test fairness. And in fact, there seem to be no recorded instances of a test being taken out of use because of a failure to meet this standard.

Cleary's landmark article ushered in an era of "culture-fair" selection research, in which a host of alternative fairness definitions sprang up. The goal of this area of study, according to Petersen and Novick (1976, p. 3), was to "eliminate cultural or racial unfairness arising from the use of tests" in selection.[3] Thorndike (1971) showed that when Cleary's definition of fairness is met, the test in question will generally be unfair to the lower-scoring group "in the sense that the proportion qualified on the test will be smaller, relative to the higher-scoring group, than the proportion that will reach any specified level of criterion performance" (p. 63). According to Thorndike's constant ratio model, fairness requires that cut scores for each group be determined such that the ratio of the proportion selected to the proportion successful on the criterion measure is the same for all groups. (As in most of the fairness definitions emerging from this era, success on the criterion is conceptualized as performance exceeding a particular cutpoint.) Some circumstances are fair according to the constant ratio model, but not the Cleary model (Petersen & Novick, 1976). Only if the predictor had perfect validity and the two groups had the same regression equation would the Cleary and Thorndike definitions agree that the predictor was fair.

Numerous other fairness definitions were advanced during the 1970s. Cole (1973) suggested that the conditional probability of selection, given success on the criterion, should be the same for all groups. Einhorn and Bass (1971) proposed that each group's cut score on the predictor should be determined such that the risk of failure on the criterion at the cut score is the same across groups. If two groups have identical regression slopes and intercepts, but different residual variances, selection will be fair according to Cleary's definition but not according to the Einhorn-Bass model. Darlington (1971) described—and dismissed—four possible formulations of culture-fairness, based on the intercorrelations of three variables—a predictor, a criterion, and a group membership variable. He proposed instead a prediction model that would require the user to specify, in quantitative terms, the value of selecting members of certain cultural groups. In effect, Darlington's model "would add a specified number of points to the scores of one subpopulation and then use the same predictor cut score" for all groups (Petersen & Novick, 1976, p. 23).

Gross and Su (1975) and Petersen (1976; see also Petersen & Novick, 1976) proposed an approach based on decision theory. In her discussion of the method, Petersen assumes that "minority" and "majority" group members are to be considered for admission based on a predictor score, and that "success" can be defined dichotomously in terms of some criterion variable. "Qualified" candidates are those who have the potential to be successful. Utility values are assigned to each possible selection decision (e.g., "accept a qualified candidate who is a minority group member"). The method can then be used to produce the predictor cut scores for minority and majority groups that maximize the expected utility.

None of the many fairness definitions put forward during the culture-fair selection era gained general acceptance. A fundamental obstacle was the lack of consensus as to whether selection procedures should incorporate social justice goals—an issue that is still unresolved. The fairness definitions that took shape during this period explicitly embodied such goals and generally required that, to promote them, cut scores on selection measures be allowed to differ across demographic groups, an idea that was controversial then and remains so today. In addition, both the specific rationales underlying the proposed approaches and the methods themselves were complex and difficult to explain. Finally, researchers pointed out that not only did the methods often produce very different conclusions (see Breland & Ironson, 1976 for a real-data example); some were not even internally consistent. Petersen and Novick (1976) showed that both Thorndike's and Cole's definitions were internally contradictory in the sense that if they were reformulated in terms of the proportions of individuals rejected, rather than the proportions selected, the resulting selection rule would, in general, differ from the rule obtained using the usual formulation. This was a troubling finding because "conceptually, it seems just as reasonable to explicate the fundamental concept of each approach by exhibiting concern for the rejected and/or unsuccessful applicant" (Petersen & Novick, 1976, p. 14).

Another dilemma that led to some discussion during this era (though it is not unique to the methods proposed at the time) is the question of which groups require special attention. For example, if we were to use Cleary's approach, for which groups would we need to obtain and compare separate regression equations? During the culture-fair selection era, several researchers made facetious suggestions about the groups that should be considered, with Cronbach (1976) asking whether those with Type O and Type A blood needed to be compared and Darlington (1976) questioning whether short people and atheists warranted special consideration.

Where did the many models for culture-fair selection lead? Despite the lack of a universally accepted technical definition of fairness, the models developed during this period spurred important thinking about ways to use data analyses to investigate issues of equity. Although they provide no definitive answers, statistical analyses, such as examinations of differential prediction, differential validity, and differential item functioning, have become widely accepted as necessary steps in evaluating test and selection fairness. However, the technical developments that arose from the culture-fair selection era had little or no impact on public views of the use of tests in selection. According to Cole (1981, p. 1067), "the large amount of technical work on bias in the last 15 years cannot be said to have significantly clarified the public controversies or to have resolved many issues associated with bias in the public mind." Hunter and Schmidt (1976, p. 1069) stated, "We feel that we have shown that any purely statistical approach to the problem of test bias is doomed to rather immediate failure." Cronbach (1976, p. 31) put it even more succinctly: "Make no mistake. The issues will not be settled by mathematical specialists."

Definitions from Philosophy

If statistics and psychometrics are unable to offer a universally acceptable formulation of fairness, what about philosophy? After all, an entire field of study within philosophy—distributive justice—addresses the manner in which societal benefits are allocated. More specifically, "the economic, political, and social frameworks that each society has . . . result in different distributions of benefits and burdens across members of the society . . . Arguments about which frameworks and/or resulting distributions are morally preferable constitute the topic of distributive justice" (Lamont & Favor, 2017).

A few attempts have been made to apply philosophical standards to selection practices in education. In their book on college admissions, *Leveling the Playing Field*, philosophers Robert Fullinwider and Judith Lichtenberg (2004) offered two governing principles for college admissions. (Here, I draw on the discussion in Zwick, 2017.) First, "other things being equal, it is desirable to enhance educational opportunities for those whose opportunities have been significantly limited" (Fullinwider & Lichtenberg, 2004, p. 11). Although this sounds eminently reasonable, it leaves key issues unresolved. What other things must be equal? Grades? Test scores? And how should limitation of opportunities be evaluated? Would a childhood illness or an unsupportive family count as a limitation? According to the second principle, "individuals should be neither helped nor hindered in their efforts at educational advancement by factors irrelevant to the legitimate goals of educational institutions" (p. 13). But here again, it is not easy to see how this principle would be applied in practice. As stated, it raises difficult questions about legitimate goals of universities. If staying afloat financially is legitimate, that suggests that using personal wealth to influence admissions decisions should be acceptable. Fullinwider and Lichtenberg dismiss this troubling possibility by stating that maintaining financial solvency is not the core mission of universities. Even if we put aside financial issues, however, the second principle falls short of providing clear guidelines for practice because of the lack of consensus about the goals of institutions of learning. Is redressing social wrongs a legitimate goal? Is fostering diversity legitimate? And if either of these goals is appropriate, what mechanisms should be used to attain them? In American society, we are far from achieving unanimity on these issues.

As part of a book chapter on philosophical perspectives on test fairness, Neil Dorans and I offered some conjectures about the stances various philosophers would take regarding college admissions practices (Zwick & Dorans, 2016). We considered the ideas of Aristotle and of two modern-day philosophers—Robert Nozick and John Rawls.

Aristotle argued that people should be given what they deserve based on their talents and accomplishments, rather than on their position in society. He also believed that the just distribution of goods depended on the purpose of the good. For example, in *Politics*, a work of political philosophy written in the fourth century BCE, Aristotle considered the distribution of flutes and concluded that "when a number of flute players are equal in their art, there is no reason why those of them who are better born should have better flutes given to them; for they will not play any better on the flute, and the superior instrument should be reserved for him who is the superior artist" (Aristotle, 2005, p. 47).

Here again, the question of the legitimate goals of universities arises. We speculated that if a university's purpose were solely to foster academic excellence, basing admissions entirely on academic criteria such as grades and admissions test scores would be acceptable, assuming that these measures were considered to be valid measures of academic talents. But is that the only purpose of a university? If a university's role includes the promotion of social goals, such as encouraging diversity or broadening educational opportunity, then factors other than narrowly defined academic skills must be considered (Zwick & Dorans, 2016; see also Sandel, 2009).

Aristotle's perspective on selection would, in any case, be considered grossly unjust by today's standards because of his defense of slavery and his belief that neither slaves nor women qualified for citizenship or any of its associated benefits. Aristotle contended that some men were well-suited to serve as slaves and that for them, slavery was "both beneficial and just" (Sandel, 2009, p. 202). That some men were masters and others slaves was "natural" in Aristotle's eyes. Like slaves, women were inferior beings, destined to be ruled by free men and ineligible for citizenship. The role of both slaves and women was to liberate free men from everyday household concerns, allowing them to participate in the important activities of the political community (Sandel, 2009; Stauffer, 2008).

Recognizing the complexities of attempting to apply ancient philosophies to today's world, we moved on to two 20th-century philosophers. We started with the free-market libertarian ideas espoused by Robert Nozick. From the libertarian perspective, inequality per se is not injustice, and Nozick (1974) specifically opposed the idea that a fair allocation of goods requires a particular pattern of distribution. Wealthy individuals may use their superior financial resources as they wish, as long as their wealth has been legitimately obtained. In the admissions context, they are free to use their wealth to buy the best test preparation courses for a candidate or even to influence a school's selection decision. In fact, according to libertarian thinking, college seats could be auctioned to the highest bidder.

Coincidentally, as I was preparing the address on which this paper is based, the "Varsity Blues" scandal broke. Federal prosecutors charged 50 people in a fraudulent scheme to purchase admission to prestigious universities. Thirty-three of those charged were parents of the applicants (Medina, Benner, & Taylor, 2019). The idea of rich families openly using their wealth to buy favorable admissions decisions for their children was not an abstraction, but a reality that was featured prominently in the headlines.

And although I had written about auction-based admissions under the assumption that it would be considered a preposterous idea, that very approach was proposed in the *Washington Post* in the wake of the scandal. The economics columnist Robert Samuelson (2019) offered the following recommendation: "Auction off some of those scarce spots. To the highest bidders go the admissions places." Samuelson noted that the proposal would apply only to schools with extremely low acceptance rates and only to a small proportion of the available slots. Also, applicants would have to be prescreened to be eligible for the auction pool. Finally, all successful bids would have to be paid. As Samuelson explained, "Assume that an applicant applies to Yale, Stanford, Harvard and MIT—and gets into all four. She picks MIT. But Mom and Dad would still have to foot the bill for Yale, Stanford and Harvard." According to Samuelson, this system would provide a morally acceptable way to force the wealthy to "pay more for their good fortune." College admission via auction has been discussed before (see Klitgaard, 1985; Sandel, 2009), but not as a serious policy proposal. Was Samuelson's suggestion facetious? In light of the impersonations, lies, and bribes of Varsity Blues, it is hard to know. The work of Daniel Golden suggests that we are perilously close to a system in which college places can be openly purchased (albeit without an auction). In *The Price of Admission* (Golden, 2006), he details the phenomenon in which certain below-par students are accepted by elite institutions as "development admits"—applicants whose families are considered to be a good source of donations. However, this practice is regarded as grossly unfair by a majority of the U.S. population. In a recent national poll conducted by *USA Today* and Suffolk University, 83% of respondents said that "it is not acceptable for students to get special treatment if their parents or relatives contribute large sums of money to a university or buy a building" (Page & Berry, 2019), thus rejecting the libertarian perspective.

In our chapter, Dorans and I also considered the complex ideas on justice and fairness presented by John Rawls, some of which are relevant to the admissions setting. Rawls believed that in general, social goods must be distributed equally; any inequalities must benefit all citizens, especially the least advantaged in society (see Wenar, 2017). According to Rawls, educational resources should be allocated "so as to improve the long-term expectation of the least favored. If this end is attained by giving more attention to the better endowed, it is permissible; otherwise not" (Rawls, 2007, p. 218). In a brief discussion of Rawls, Cronbach offered an interpretation of how this principle might translate into practice, suggesting that "the practice of competitive admissions to medical schools, which lowers the chance of children born into poor families to become physicians, might be justified on the ground that this raises the standard of health care for the great mass of the public" (Cronbach, 1976, p. 36).

More controversial than Rawls's toleration of certain types of inequality are his views on "desert." Rawls believed that we do not deserve to be rewarded for our talents any more than we deserve to be rewarded for wealth or social position. Talents, after all, are merely the result of a "natural lottery." Taking this idea even further, Rawls argued that we do not deserve to be rewarded for effort either, because effort, too, is ultimately a function of luck and opportunity: "Even the willingness to make an effort, to try, and so to be deserving in the ordinary sense is itself dependent upon happy family and social circumstances" (Rawls, 2007, p. 216). Clearly, Rawls would not have endorsed the current glorification of grit as an indicator of academic potential. As Dorans and I put it, "neither a perfect GPA, nor an Olympic medal, nor a year volunteering in an inner-city soup kitchen would automatically 'merit' admission to UC Berkeley, from a Rawlsian perspective" (Zwick & Dorans, 2016, p. 276). It appears that Rawls's principles would disallow most existing admissions methods, which do, in fact, reward various combinations of talent and effort.

Dorans and I concluded our foray into philosophy as follows:

> Can our three, or any, philosophical perspectives provide conclusive answers to our fairness dilemmas? We don't think they can ... However, [they] can provide us with alternate lenses through which to view assessment fairness issues and can at least encourage us to ask the right questions. (Zwick & Dorans, 2016, p. 279)

Yet another examination of higher education admissions from a philosophical viewpoint was conducted by Meyer (2013). He too examined the work of Aristotle and Rawls, and also considered libertarianism, utilitarianism, and the work of Emmanuel Kant. His conclusion about the usefulness of formal theories on fairness and ethics: "While these lines of reasoning generate ideas and insights ... they are indeterminate with respect to the kind of rules or policies that can lead to more just institutional arrangements" (Meyer, 2013, p. 16).

The upshot, then, is that neither statistics nor philosophy has yielded a simple, satisfying definition of fairness in measurement and selection. More significantly, neither statistical procedures nor philosophical pronouncements have been helpful in addressing public concerns about selection fairness. For most people, fairness is personal: Their views on testing are based on their own experiences or those of their friends and family, and their concerns cannot be addressed through explanations of overprediction, treatises on Aristotelian views of merit, or discussions of Rawlsian concepts of deservingness.

Public Perspectives

It is useful to start by recognizing that selection decisions are challenging by nature. Every method for allocating prized goods, such as college seats or jobs, will be viewed as unfair by some individual or entity, and this is true even when tests are not involved. Consider the case of the lottery that was used to select a portion of the freshman class of 1970 at the University of Illinois. More than 800 students, some of whom were at the top of their high school classes, were rejected as a result. Here is what the mother of one of those students told the *Chicago Tribune*:

> He doesn't have long hair, doesn't smoke, and doesn't drink. He gave up playing football so he could go to work ... He has an Illinois state scholarship, and has made especially good college test scores. We didn't know that there would be a lottery or we may have started looking for another school in the first place. (Buck, 1969, p. 5)

In fact, the public outrage following the lottery was so extreme that the lottery results were rescinded and all the rejected students were accepted (U. of I. opens doors for 839 barred by admissions lottery, 1969). The public reaction was a clear demonstration that, even when test scores are explicitly excluded from the selection process, admissions decisions can be controversial.

More familiar are the kinds of challenges that involve the role of tests in selection, as exemplified by the statements below. The first two are from authors of the book, *The Myths of Standardized Tests: Why They Don't Tell You What You Think They Do* (Harris, Smith, & Harris, 2011). The third is from the "Dear Therapist" column in *The Atlantic*.

> My youngest daughter ... attended Indiana University and graduated with a degree in kinesiology. She wanted to go to graduate school to study physical therapy but could not gain admittance because of low GRE scores. But she believed in herself and knew that the test scores weren't a true measure of her ability to succeed in graduate school, and, today, she has completed a PhD. (Harris, Smith, & Harris, 2011, p. 49)

> My wife got violently sick before she took the SAT, and there seemed to be no physical cause. What's more, it's not something she has outgrown ... When her maturity, work ethic, and organizational skill ... led [her] company to ask her to take their management training, she took the workshops and studied the material. I know she knew it, because I quizzed her on it. When she sat for the test, she froze, as always ... The company said it couldn't do anything for her. The test scores, after all, were objective. (Harris, Smith, & Harris, 2011, p. 68)

> My son is in the middle of the college-application process. He has very good grades and very good SAT and ACT scores; he is an Eagle Scout and a captain of the cross-country team ... According to all of the statistics and reports, he should be accepted at Ivy League schools, but he has not been. How do you explain to a bright, eager boy that the system is rigged against him? (Gottlieb, 2019)

Before considering possible responses, let's review the essence of each of these challenges to testing and selection practices. The first scenario concerns a woman who was not admitted to her graduate school of choice; her father (who wrote the quoted text) attributes this to her low GRE scores. The second vignette describes a woman who had severe anxiety when she took the SAT and continues to have test anxiety as an adult. According to her husband (who wrote the text), her test scores are therefore poor reflections of her abilities. The third vignette, written by the mother of a college applicant, is a complaint that, despite his many qualifications, the writer's son was nevertheless rejected by Ivy League Schools.[4]

Addressing Public Challenges

How can those of us in the measurement field respond in a useful way to these sorts of challenges to testing and selection practices? First, we can assure that test users and test takers are aware of existing standards. More broadly, we can continue to take a stand against misuses and advocate for fair use of testing and selection practices. Finally, we can communicate candidly with critics about testing and selection issues. I discuss each of these three types of responses in more detail in the following sections.

Assuring that test users and test takers are aware of standards. As a field, we have standards in place that are relevant to the GRE-taking kinesiologist, the anxious SAT-taker, and the Eagle Scout rejected by the Ivy League. For example, we have a standard about the use of multiple measures: "In educational settings, a decision or characterization that will have a major impact on a student should take into consideration not just scores from a single test but other relevant information" (AERA, APA, & NCME, 2014, p. 198). This standard is relevant to all three vignettes. Perhaps if the test users—the institutions using tests for selection purposes—had been aware of this standard, different selection decisions might have been reached. Depending on the particulars of the selection processes, this standard might have also provided the basis for a challenge by the rejected applicants.

Consistent with the standard on the importance of multiple measures, all reputable testing companies recommend against over-reliance on tests. Of particular relevance to the GRE-taking kinesiologist is the following excerpt from the *GRE Guide to the Use of Scores:*

> It is important that programs not over-rely on GRE scores, and never use GRE scores as the sole criteria for "cut scores" ... Rather, faculty reviewers might consider adopting practices that put GRE scores into the appropriate context in relation to the other elements an institution might require that candidates submit (Educational Testing Service, 2018, p. 5).

As a field, we also have principles about the need for transparency of testing practices. According to the *Standards*, test takers should be informed in advance about "the test, the testing process, the intended test use, test scoring criteria, [and] testing policy" (AERA, APA, & NCME, 2014, p. 134), especially when stakes are high (p. 131). I have argued that in the admissions context, the entire candidate evaluation procedure should be regarded as the "test" and that much more information should be provided to applicants about admissions processes and criteria (Zwick, 2017). Maintaining transparency in admissions is particularly important because sociological research tells us that the less clear and specific the admissions criteria, the greater the disadvantage suffered by those not already knowledgeable about academic culture (Zwick, 2016). Perhaps if the Eagle Scout and his mother knew more about the complexities of college admissions decisions, they would have been less distressed by the outcome (a point that was made in the published response to this letter).

Advocating for fair use of testing and selection practices and taking a stand against misuses. One of the ways that the National Council on Measurement in Education has worked to promote good testing practices is to issue position statements, available at ncme.org. So far, NCME has issued statements on student participation in state assessment, theories of action for testing programs (descriptions of intended outcomes and how they will be achieved), K-12 classroom assessment, test security, and the use of admissions testing for accountability purposes. Statements are currently being drafted on the design of balanced testing programs and on the assessment of English language learners. As measurement specialists, we can also promote good practices and identify misuses in our professional work as researchers, consultants, members of advisory committees, and authors, and in our informal water-cooler and cocktail-party communications.

Communicating candidly with critics about testing and selection issues. As Worrell (2016, p. 291) notes,

> [i]t is incumbent upon ... measurement professionals to not only design tests that yield reliable and fair scores from which valid inferences can be generated but also to proactively communicate with and educate the public and policymakers about testing, test construction, test utility, and *test fairness*.

How can we communicate more effectively with test critics? First, we can make an effort to find common ground. We can acknowledge that even the best tests are fallible (i.e., imperfectly reliable), that some tests are just plain bad, that tests are sometimes misused, and that there is such thing as too much testing. More generally, we should strive to avoid defensiveness about testing and its role in selection.

Conclusions

In industrialized societies, selection decisions that involve standardized tests play a substantial role in many people's lives. Despite this, no canonical definition of fairness in measurement and selection has emerged from the fields of educational and psychological testing. The closest we have come is the definition offered by Cleary in 1968, which declares a test to be biased if the regression of the criterion of interest on the test score differs among key demographic groups. This definition has been subjected to extensive critiques, however, and appears only in diluted form in the current *Standards*. The alternative statistically based definitions proposed during the 1970s to promote culture-fair selection have not fared any better. Although they led to fruitful and interesting discussions, none of these models are used widely, if at all, today. The writings of philosophers, ancient and modern, provide some intriguing perspectives on selection fairness, but also fail to yield a simple rule. And neither the statistical nor the philosophical definitions are well aligned with the concerns of the public, which often stem from personal experiences with selection decisions. Experiences with lottery admissions show that even when tests are not involved, selection processes can be controversial and decisions can be hard to accept. How can those of us in the measurement field respond in a useful way to challenges to testing and selection practices? We can do our best to assure that test users and test takers are aware of existing standards, take a stand against misuses, advocate for fair use of testing and selection practices, communicate candidly with critics about testing and selection issues, and acknowledge problems where they exist. Our role as measurement specialists should not be to defend tests at all costs. Instead, we should be the judicious evaluators of tests and their applications.

Notes

[1] The degree of overprediction was greater when high school grades were included as predictors. A likely explanation is that grading was generally less stringent at the high schools attended by the Black students than at the schools attended by the White students, so that a given high school grade corresponded to a lower level of performance for Black than for White students.

[2] It is interesting that Einhorn and Bass (1971) attribute the equal-regressions definition of fairness not to Cleary but to Anastasi (1968). Anastasi (1968, pp. 559–561), in turn, references Cleary (1966), which is an early version of Cleary (1968), along with other roughly concurrent studies. Like Einhorn and Bass, Jensen (1980) emphasized the importance of possible differences across groups in the size of the residual variance. His definition is framed in terms of a hypothetical test with perfect reliability. Such a test is said to be a biased predictor if the regression systems for major and minor groups differ in terms of slopes, intercepts, or residual variances. Jensen noted that Cleary's definition of unbiased prediction, which makes no mention of residual variances, is less stringent than his own. It may seem ironic that an author better

known for his inflammatory views on racial differences in intelligence emerges here as a proponent of rigorous criteria for test fairness.

[3] Petersen and Novick (1976) define the regression model for evaluating test fairness, which they attribute to Cleary, in a way that differs somewhat from Cleary's own definition. Although Cleary states that the absence of bias requires that a single regression model fit the data for all groups, Petersen and Novick also consider it acceptable if regressions differ, as long as the cut scores on the predictor are appropriately adjusted. They state that, according to the regression model, "if two applicants are being considered for one post, then that applicant having the highest predicted performance would be selected with prediction being made on the basis of subpopulation regression" (p. 5).

[4] The letter includes the dubious statement, "He is also white, male, and upper-middle-class—and that is the problem." It seems that the writer believes that race, gender, and class were factors in her son's unsuccessful application efforts. This claim is undercut by her own words contrasting her son's experience with the much smoother experience of his twin brother, who is presumably also white, male, and upper-middle-class. I have chosen to focus on other aspects of the letter.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Anastasi, A. (1968). *Psychological testing* (3rd ed.). New York, NY: Macmillan.

Aristotle. (2005). *Politics*. Stilwell, KS: Digireads.

Breland, H. M., & Ironson, G. H. (1976). DeFunis reconsidered: A comparative analysis of alternative admissions strategies. *Journal of Educational Measurement*, *13*, 89–99.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Buck, T. (1969, December 6). Mother irked by son's loss in U. I. lottery. Chicago Tribune, p. 5.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport, CT: American Council on Education/Praeger.

Cleary, T. A. (1966). Test bias: Validity of the Scholastic Aptitude Test for Negro and white students in integrated colleges (Research Bulletin 66-31). Princeton, NJ: Educational Testing Service.

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

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement*, *10*, 237–255.

Cole, N. S. (1981). Bias in testing. *American Psychologist*, *36*, 1067–1077.

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

Culpepper, S. A., Aguinis, H., Kern, J. L., & Millsap, R. (2019). High-stakes testing case study: A latent variable approach for assessing measurement and prediction invariance. *Psychometrika*, *84*, 285–309.

Darlington, R. B. (1971). Another look at "cultural fairness". *Journal of Educational Measurement*, *8*, 71–82.

Darlington, R. B. (1976). A defense of "rational" personnel selection, and two new methods. *Journal of Educational Measurement*, *13*, 43–52.

Educational Testing Service. (2018). GRE guide to the use of scores, 2018–19. Princeton, NJ: Author.

Einhorn, H. J., & Bass, A. R. (1971). Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, *75*, 261–269.

Flaugher, R. L. (1970). Testing practices, minority groups, and higher education: A review and discussion of the research (Research Bulletin 70-41). Princeton, NJ: Educational Testing Service.

Fullinwider, R. K., & Lichtenberg, J. (2004). *Leveling the playing field: Justice, politics, and college admissions*. Lanham, MD: Rowman & Littlefield.

Golden, D. (2006). *The price of admission*. New York, NY: Crown Publishers.

Gottlieb, L. (2019, February 18). Dear therapist: I'm worried the college-admissions process is rigged against my son. The Atlantic. Retrieved from https://www.theatlantic.com/family/archive/2019/02/im-worried-my-son-wont-get-good-college/582979/

Gross, A. L., & Su, W. (1975). Defining a "fair" or "unbiased" selection model: A question of utilities. *Journal of Applied Psychology*, *60*, 345–351.

Harris, P., Smith, B. M., & Harris, J. (2011). *The myths of standardized tests: Why they don't tell you what you think they do*. Lanham, MD: Rowman & Littlefield.

Humphreys, L. G. (1952). Individual differences. *Annual Review of Psychology*, *3*, 131–150.

Hunter, J. E., & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin*, *83*, 1053–1071.

Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: The Free Press

Klitgaard, R. E. (1985). *Choosing elites*. New York, NY: Basic Books.

Lamont, J., & Favor, C. (2017). Distributive justice. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2017). Stanford, CA: Stanford University. Retrieved from https://plato.stanford.edu/archives/win2017/entries/justice-distributive/

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

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement*, *21*, 33–47.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, *8*, 1–4.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). Differential validity and prediction of the SAT (College Board Research Report No. 2008-4). New York, NY: The College Board.

Medina, J., Benner, K., & Taylor, K. (2019, March 12). Actresses, business leaders and other wealthy parents charged in U.S. college entry fraud. *New York Times*. Retrieved from https://www.nytimes.com/2019/03/12/us/college-admissions-cheating-scandal.html

Meyer, H.-D. (2013). Reasoning about fairness in access to higher education: Common sense, normative, and institutional perspectives. In H.-D. Meyer, E. P. St. John, M. Chankseliani, & L. Uribe (Eds.), *Fairness in access to a higher education in a global perspective* (pp. 15–40). Rotterdam, The Netherlands: Sense Publishers.

Nozick, R. (1974). *Anarchy, state and utopia*. New York, NY: Basic Books.

Page, S. & Berry, D. B. (2019, March 20). Poll: Americans say even the legal breaks for college admission rig the system. *USA Today*. Retrieved from https://www.usatoday.com/story/news/politics/2019/03/20/poll-college-admissions-unfair-favor-wealthy/3212228002/

Pennock-Román, M. (1994). College major and gender differences in the prediction of college grades (College Board Report 94-2). New York, NY: College Entrance Examination Board.

Petersen, N. S. (1976). An expected utility model for "optimal" selection. *Journal of Educational Statistics*, *1*, 333–358.

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

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). Student group differences in predicting college grades: Sex, language, and ethnic groups (College Board Report 93-1). New York, NY: College Entrance Examination Board.

Rawls, J. (2007). Rawls: Justice as fairness. In M. J. Sandel (Ed.), *Justice: A reader* (pp. 203–221). New York, NY: Oxford University Press.

Samuelson, R. J. (2019, March 19). Here's how to fix the college admissions system. Warning: You might hate this. *Washington Post*.

Retrieved from https://www.washingtonpost.com/opinions/heres-how-to-fix-the-college-admissions-system-warning-you-might-hate-this/2019/03/19/78eb40c4-4a66-11e9-9663-00ac73f49662_story.html?noredirect=on

Sandel, M. J. (2009). *Justice: What's the right thing to do?* New York, NY: Farrar, Straus, & Giroux.

Schmidt, F. L., & Hunter, J. E. (1974). Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. *American Psychologist*, *29*, 1–8.

Stauffer, D. J. (2008). Aristotle's account of the subjection of women. *The Journal of Politics*, *70*, 929–941.

Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in predictions of college grades from Scholastic Aptitude Test scores. *Journal of Educational Psychology*, *85*, 710–718.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, *8*, 63–70.

U. of I. opens doors for 839 barred by admissions lottery. (1969, December 13). U. of I. opens doors for 839 barred by admissions lottery. Chicago Tribune. Retrieved from http://archives.chicagotribune.com/1969/12/13/page/1/article/u-of-i-opens-doors-for-839-barred-by-admissions-lottery

Wenar, L. (2017). John Rawls. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017). Stanford, CA: Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2017/entries/rawls/

Worrell, F. C. (2016). Commentary on perspectives on fair assessment. In N. J. Dorans & L. Cook, (Eds.), *Fairness in educational assessment and measurement* (pp. 283–293). New York, NY: Routledge.

Young, J. W. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In Zwick, R. (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 289–301). New York, NY: RoutledgeFalmer.

Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. New York, NY: RoutledgeFalmer.

Zwick, R. (2016, January 17). Transparency in college admissions is key to a fair policy on race. Chronicle of Higher Education. Retrieved from http://chronicle.com/article/Transparency-in-College/234949

Zwick, R. (2017). *Who gets in? Strategies for fair and effective college admissions*. Cambridge, MA: Harvard University Press.

Zwick, R., & Dorans, N. J. (2016). Philosophical perspectives on fairness in educational assessment. In N. J. Dorans & L. Cook, (Eds.), *Fairness in educational assessment and measurement* (pp. 267–281). New York, NY: Routledge.

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*, 101–121.