Anticipating Likely Consequences of Lottery-Based Affirmative Action*

Bernard Grofman, *University of California, Irvine*
Samuel Merrill, *Wilkes University*

Objectives. To better understand the consequences of whole or partial reliance on test scores as a screening mechanism for college or university admissions or for job placement or promotion. *Methods.* We introduce a simple hurdles/threshold model to study one particular mechanism that has been recently proposed to generate a "compromise" between race-normed or gender-normed standards for acceptance, on the one hand, and the decision to select those with the highest score regardless of race or gender, on the other—a method we call *lottery-based rules with minimum thresholds of acceptance*. *Results.* We show the factors that determine how close acceptance rates for the disadvantaged group under lottery-based methods will be to the acceptance rates under either race-normal or pretest-score-based mechanisms. *Conclusions.* We argue that the likely consequences of using this method are not nearly as attractive as they might first appear to reformers.

We can certainly find examples of "objective" tests as a choice mechanism in preindustrialized nations (Mandarin China is an example that immediately leaps to mind), but modern industrialized societies have elevated testing as a mode of selection to heights never before seen. The United States, in particular, can readily be characterized as a "test-taking society." Many important decisions (as to admissions, hiring, and promotion) are reached, at least in large part, on the basis of test results. This fact can have enormous implications for equality within and across groups categorized by attributes such as race or gender.

The role of testing in job hiring and promotion, and in all levels of education from transfer to magnet high schools to admission to prestigious law schools and medical schools, has been (and continues to be) the subject of much political controversy, and of legislation and litigation. Various modes of affirmative action have been proposed to compensate for the fact that there are substantial differences in mean performance by race on most

*Address correspondence to Bernard Grofman, School of Social Sciences, University of California, Irvine, CA 92697 ⟨bgrofman@uci.edu⟩. The first-named author will share all data and coding information with those wishing to replicate the study. We are indebted to Clover Behrend-Gethard, Dorothy Green, and Chau Tran for library assistance. This research was partially supported by the Program in Methodology, Measurement and Statistics, National Science Foundation, via SBR #97-30578 (to Bernard Grofman and Anthony Marley). The listing of authors is alphabetical.

SOCIAL SCIENCE QUARTERLY, Volume 85, Number 5, December 2004

standardized paper-and-pencil tests and in school grade distributions as well (see, e.g., Jencks and Phillips, 1998). Methods proposed to mitigate the consequences of such differences for admissions, hirings, and promotions include racial quotas, different test score or grade point "cutoffs" for different groups, bonus points for "strivers" who overachieve relative to their group's average performance, lottery methods, methods that use class background as a substitute for race,[1] and methods designed to provide geographic balance with the aim of also generating more racial and ethnic balance in light of patterns of residential segregation (e.g., rules that automatically admit to state universities students who are in the top k% of their high school class regardless of where they stand relative to the statewide applicant pool as a whole).[2] Recently, there have been highly publicized referenda on the use of racial or gender criteria by state governments that have considerably affected affirmative action policies in California and Washington (including admission policies for the state universities). In 2003 there were two major Supreme Court decisions revisiting the issues about the use of race in admissions decisions in higher education that lay at the heart of *Regents of the University of California v. Bakke*, 438 U.S. 265 (1978).[3]

It is the group-related implications of testing that will be the focus of this article. There is an ongoing debate on the power of tests (e.g., the SAT, ACT, LSAT, GRE) to predict school performance and subsequent life success, and a similar debate about the predictive power for job performance of various paper-and-pencil and other tests used to help determine hiring or promotion decisions. In this article we will not attempt to address this controversy about the predictive power of tests for future success.[4] Nor will we seek to contribute to the ongoing debate why group differences in test scores exist/persist (see, e.g., Jencks and Phillips, 1998). Instead, we simply focus on the racial and ethnic and gender consequences for admissions or hiring or promotion decisions of using test scores as a key screening device *in situations where we may take the group means and standard deviations on the test in question to be known*. In particular, we develop a model to specify the

[1]See especially Kahlenberg (1996).

[2]For a very useful general discussion, see Zwick (2002).

[3]The cases are *Grutter v. Bollinger*, 539 U.S. 982 (2003), dealing with law school admissions to the University of Michigan, and *Gratz v. Bollinger*, 538 U.S. 959 (2003), dealing with undergraduate admissions at the same university.

[4]Courts have been skeptical of test use when it cannot either be directly linked to job performance or shown to be a "business necessity" in that it reduces arbitrariness and the potential for bias in employment decisions (Kadane and Mitchell, 2001). Many civil rights advocates have expressed the belief that tests are often used merely to provide the "appearance" of fairness, because it will be known in advance that lower test performance can be expected from members of "undesired" minorities. On the other hand, Spence (1974) notes that even when a test is not directly related to the skills needed for job performance, test evidence may be used by employers for screening purposes as a signal of general competence and/or of willingness to invest in human capital. Of course, sometimes testing is discriminatorily administered, as was, for example, often the case with literacy tests in the American south (see, e.g., Alt, 1994).

implications for group-specific acceptance rates as a function of the test-cutoff levels that are used, and of whether the final decision to accept/promote is based solely on test scores or also on some type of lottery-like consideration. Thus our models have six key variables: the means and standard deviations of each group, the threshold, and the specific mechanism.[5]

Of course, we recognize that the desirability of any testing regime is contingent on the value of the test as a predictor of future success. Nonetheless, the empirical implications of alternative testing regimes for race-specific (or gender-specific) *acceptance rates* can be investigated without addressing the predictive value of particular tests, and without debating, in the abstract, philosophic issues having to do with affirmative action.[6]

Consider two groups that differ in either their mean or standard deviations on some attribute, which we may take to be a test score. For values of the attribute that are approximately normally distributed, it is well known among statisticians that if we impose some threshold such that only group members who score above that threshold will be chosen, then, *ceteris paribus*: (1) for unequal group means and identical standard deviations, the higher the threshold, the greater will be the acceptance rate of the group with higher mean relative to that of the group with lower mean; and (2) for identical group means and unequal standard deviations, the higher the threshold (above the common mean), the greater will be the acceptance rate of the group with higher standard deviation relative to that of the group with lower standard deviation (see, e.g., Paulos, 1995; Berger, Wang, and Monahan, 1998).

These simple statistical insights (based on the properties of the normal distribution and related unimodal distributions) have important implications in a variety of applications. For example, the interaction between thresholds and acceptance rates can help explain why we may observe very low proportions of members of some groups in situations where the selection process chooses only those who are rather far out on the tails of distributions (Crow, 2002:85). As we shall see, the magnitude of these "tails of distributions" effects can be rather astonishing, even to people who would see themselves as generally familiar with the properties of the normal distribution. Also, and even more importantly, despite this "tails of distribution" phenomenon being well known to mathematicians (Poulos, 1995) and highlighted by some students of educational policy (see, e.g., Kane, 1998:435; Hedges and Nowell, 1999:125–30; Zwick, 2002), it is almost entirely neglected in the legal discussion of affirmative action issues.

The purpose of this article is to make use of insights about "tails of distributions" in the study of alternative modes of affirmative action. We begin with an elementary exposition of the statistical logic underlying the

[5]However, we can usually simplify the model by considering ratios (or differences) of the mean and standard deviation parameters (see below).

[6]For a insightful analytic discussion about competing ideas of equality, see Rae et al. (1981).

basic threshold model. Although the approach we offer is a very general one that can be applied to analyze the acceptance-rate consequences of any of a wide variety of affirmative action methods, the heart of our article focuses on the implications for acceptance rates of a recently proposed method to implement affirmative action, what we call *lottery-based rules with minimum thresholds of acceptance*. Such rules set a minimum threshold for acceptance and use a lottery to choose among applicants who score above that threshold to achieve a fixed number of "winning" candidates. In the United States, such rules have recently been used in the context of affirmative action for (magnet) high school admissions in San Francisco and have been advocated as desirable for implementation elsewhere (Cohen, 1994; Guinier, 1996; cf. Berger, Wang, and Monahan, 1998) as a compromise/way out of the long-standing debate between advocates of race-normed standards and advocates of so-called merit (i.e., highest test-score based) hiring.[7]

With some appropriate threshold chosen, in terms of descriptive representation, we show that lottery-based rules do produce intermediate outcomes between race-normed or gender-normed standards for acceptance, on the one hand, and the decision to simply select those with the highest score regardless of race or gender, on the other. Thus, these methods can contribute to diversity. We present some mathematical insights into the question of which of these poles the outcomes of lottery-based rules are most likely to resemble. We also look at the consequences for lotteries in terms of mean test scores of the successful candidates. Here, when we use the same threshold for the lottery cutoff as we did for the least advantaged of the groups whose scores were normed to equalize success rates, lotteries must, of necessity, yield acceptance results for this group that lie below those of *both* the other procedures. We look at GPA by major (at a public university) and SAT scores by race (for the nation as a whole) to investigate how lottery-based rules and related methods might operate in the real world.

[7]We must be careful not to confuse test scores with "merit," since test scores may have only very limited predictive validity for job performance or of lifetime skill learning and retention. Moreover, even tests that might appear on their face to be neutral (i.e., race-blind and gender-blind) measures of performance (such as class ranks based on anonymous grading of exams) may nonetheless conceal important disparities. In particular, there may be differences across groups in the relative accuracy of any given test as a predictor of future performance. In this context, Guinier, Fine, and Balin (1997) show that men and women at the University of Pennsylvania Law School did not differ significantly on LSAT scores, college grade-point average, or a measure of quality of undergraduate institution, and, in general, Penn students exhibited a very narrow range on these indicators. Yet, even after LSAT scores and undergraduate grades are controlled for, race and gender remain statistically significant predictors of the grade performance of students at the University of Pennsylvania Law School, with African Americans and women scoring lower than their white and male counterparts—albeit race and gender were not nearly as powerful predictors of GPA at Penn as individual LSAT scores (see Guinier, Fine, and Balin, 1997:124–26, n.73). Guinier, Fine, and Balin argue that disparate outcomes in law-school performance between white males and other groups arise from an environment that devalues the perspectives of women or persons of color and/or emphasizes adversarial modes of interaction that help reinforce a hostile climate for women and minorities. Of course, other explanations are possible.

The Basic Hurdles/Threshold Model

In this article we offer a *single hurdle, variable threshold model* in which we consider a selection process whereby a proportion of all those who apply for a given position are accepted based on their score on some exam (e.g., only those with the highest k percent of scores are chosen; or a lottery is held among only those who have scored above some threshold). Consider n subgroups ($i = 1, \ldots, n$) where the distribution of scores on some attribute A of the ith subgroup is characterized by a mean a_i and standard deviation s_i. Let the passing score/threshold value be T. We use the expression *succeed*, that is, equal or exceed the threshold T, generically to mean "be accepted," "succeed," "pass the test," "make the positive choice," and so forth. We wish to consider what proportion of each subgroup will succeed as a function of a_i, s_i, and T.

To make the discussion easier to follow so as to avoid losing the reader in unnecessary technicalities, we will provide an informal discussion of our mathematical results in the text, with a number of illustrative tables and figures. The results we give are based on normal distributions, as are the illustrative figures and charts. However, many of our basic results go through for a very general class of probability distributions, with no particular requirements on symmetry of the distribution needed.[8] Although the results we give are based on simple statistical intuitions about the tails of distributions that are well known in the theory of reliability and survival analysis (see, e.g., Barlow and Proschan, 1965; Gross and Clark, 1975; Kalbfleisch and Prentice, 1980), the notion of threshold effects is not, as far as we are aware, part of any of the standard introductory social science statistics textbook discussions of the key properties of normal (or related) distributions. Moreover, even though the basic statistical ideas are well known to statisticians, the specific extensions of the basic threshold model to rules that combine lottery rules with minimum test-score-based thresholds appear to be original with the present authors.[9]

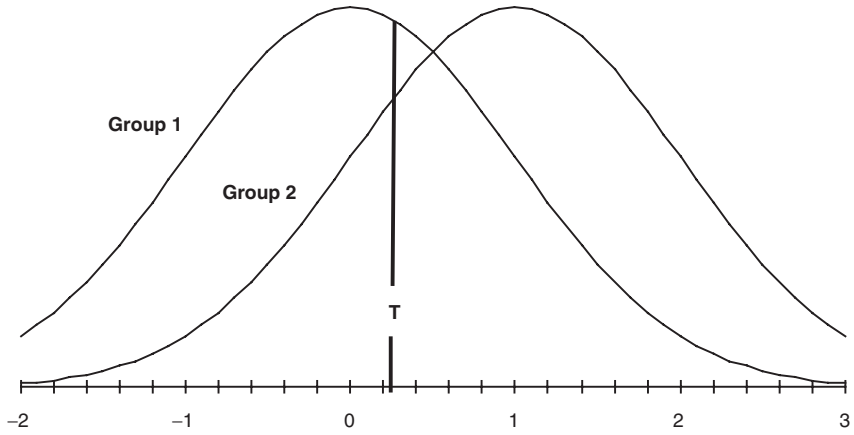
For illustrative purposes, and without great loss of generality, let us consider the case where $n = 2$. We may think of the two groups as men and women, or whites and African Americans, or Hispanics and non-Hispanics, or college graduates and noncollege graduates, and so forth. For convenience, let $0 \leq a_1 < a_2$. For a given value of T, Figure 1 allows us to visualize the *acceptance ratio* between the two groups, that is, the ratio of the proportion of members of Group 2 who succeed to the proportion of members Group 1, which is the ratio of the areas to right of the threshold for each of the two distributions. The higher this ratio, the more are Group 2 members

[8]A formal statement of the theorems we rely on (and proofs thereof) are available on request from the authors. In other work we also extend the one-hurdle model presented here to consider the implications of multiple hurdles.

[9]For reasons of space we will not attempt to discuss technical differences between our models and those of other authors such as Berger, Wang, and Monanhan (1998).

FIGURE 1

Distribution of Abilities of Groups 1 and 2 Relative to Threshold, *T*



advantaged by their test-taking abilities relative to those of Group 1 members. For convenience we will usually refer to the acceptance ratio as the *success ratio*.

Although we could analogously define an *acceptance rate difference function,* because legal standards for discrimination have usually been couched in ratio terms (see, e.g., Meier, Sacks, and Zabell, 1984), we focus on the acceptance ratio function in what follows. Let us first consider what happens to the success ratio when the two groups have different means but identical standard deviations; and when the two groups have identical means but different standard deviations.

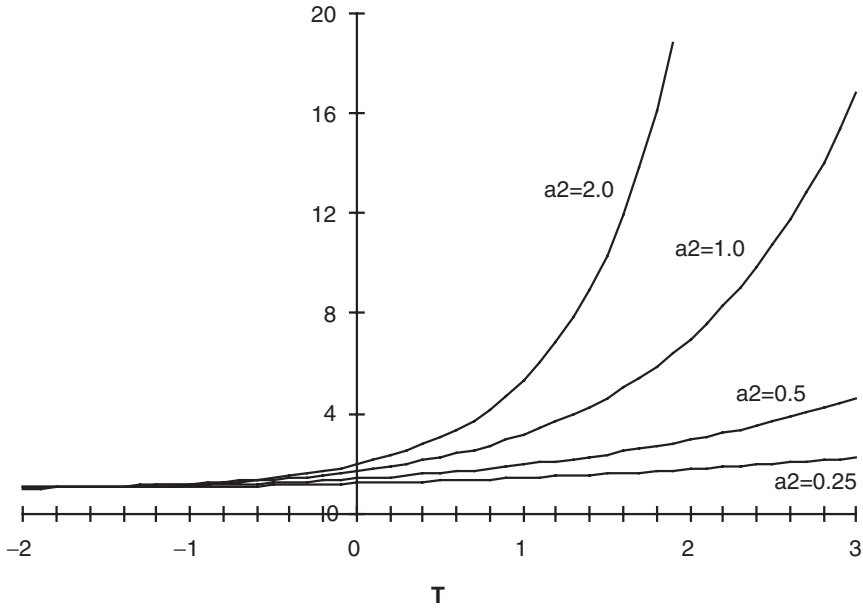
Different Means But Identical Standard Deviations

For the case of equal standard deviations and unequal means, when the hurdle is a low one relative to both the group means, proportions of each of the two groups that succeed are similar. Thus, the success ratio will be close to one when the hurdle is low. However, as the magnitude of the hurdle increases, then even modest differences in group means can make substantial differences in the proportion of each group that passes the test. When the hurdle is very high, that is, far enough out "on the tails" of each distribution we will find that representation of the least competent group is minuscule or zero, that is, the success ratio will rapidly head toward infinity for a high threshold. We illustrate this point graphically in Figure 2, showing the ratio for $s_1 = s_2 = 1$, $a_1 = 0$, and selected values of a_2, beginning at 0.25 and incrementing by a factor of 2.

If the common value, s, of s_1 and s_2 is allowed to vary, the success ratio decreases with increasing s (at least for $T > a_1$). We illustrate this point

FIGURE 2

Success Ratio as a Function of T for $s_1 = s_2 = 1$, $a_1 = 0$, and Selected Values of a_2



graphically in Figure 3, showing the ratio for $a_1 = 0$ and $a_2 = 1$ for various values of s, beginning at 0.25 and incrementing by a factor of 2.

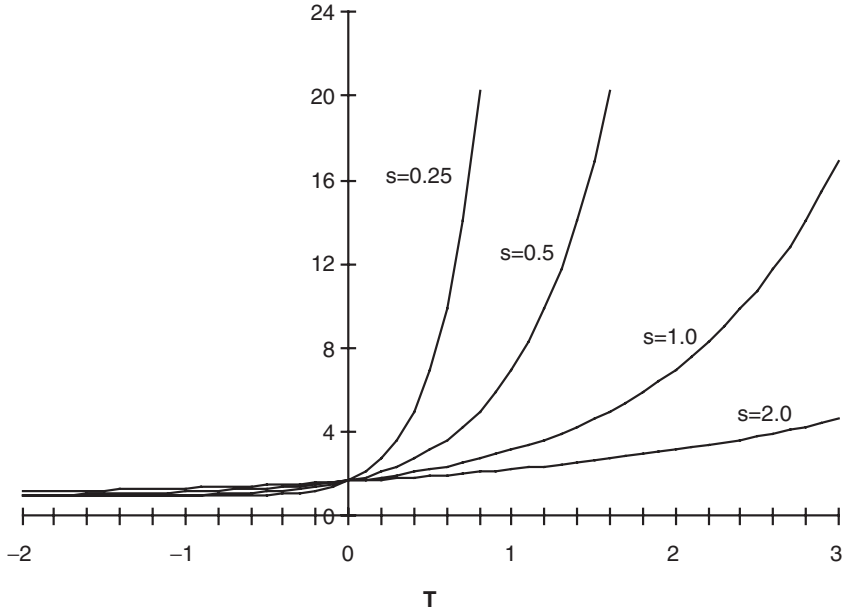
In general, the fewer who succeed, the more disproportionately will those few come from the more advantaged group, since lack of success is associated with higher hurdles.

We can make our intuitions about threshold effects even clearer by shifting from hypothetical data to real data. Imagine that a university is deciding which students to award some prize such as membership in a campuswide honor society based purely on grade-point average. Will roughly the same *proportion* of students in each major receive membership in the honor society?

We have deliberately chosen to provide our first example as one in which students are grouped by a "neutral" category like academic major rather than, say, by race or gender. That way the reader will not have to fight through ideological prejudices (either right or left) to see how statistical effects based on thresholds are generating what may be perceived of as "inequitable" treatments of members of two groups. Also, our first example, honor society membership, is one in which test scores (here grade-point averages) are *synonymous* with the performance that is being used to allocate the scarce resources in question. We have deliberately chosen to start with an example that has this property because we have discovered that, when there is a disputed link between scores on a given test and subsequent job

FIGURE 3

Success Ratio as a Function of T for Selected Values of s for Fixed $a1 = 0$ and $a_2 = 1$



performance or academic performance, it can be hard for some readers to separate out their reactions to the mathematical results about the consequences of a particular testing regime from their views about the inherent unfairness of an "unreliable" test being used as the basis of important allocational decisions.

Consider the data in Table 1 (real GPA data by major for freshmen from an unidentified public university). As we may imagine from the results we have given, the proportion of each major that will receive membership in the honor society depends not just on differences across majors in GPA distribution, but on the threshold that is used.

Let us take engineering majors as Group 1 and humanities majors as Group 2. These two sets of majors have similar standard deviations, but a moderate difference in average GPA. If the threshold is 3.20, then the success ratio is 2.08; if the threshold is as high as 3.80, the success ratio is 2.77, that is, Group 2 members *are nearly three times more likely* to meet the criterion.

Thus, uniform thresholds that are quite realistic ones for the proposed purpose of selecting members of, say, a campus honor society can have what are likely to be unanticipated consequences in terms of the likelihood that students in different majors will be admitted to the honors society. This

TABLE 1

1993 Freshmen GPA at a Major Public University

Academic Unit Code	Academic Unit Name	*N*	Average GPA	*SD* GPA
8	Unaffiliated	520	2.64	0.54
55	Biological sciences	466	2.62	0.64
57	Fine arts	80	2.91	0.56
60	Humanities	152	2.75	0.59
62	Physical sciences	187	2.54	0.70
65	Social sciences	331	2.60	0.66
77	Engineering	196	2.43	0.62
95	Computer science	116	2.46	0.78
97	Social ecology	209	2.70	0.55

empirical example, using real data, shows that differences in success rates of a nontrivial magnitude can arise from what appear to be relatively minor differences in *means* between groups. We should also note that the data are roughly normal, so that the model's simplifying assumptions are appropriate ones.

Identical Means But Different Standard Deviations

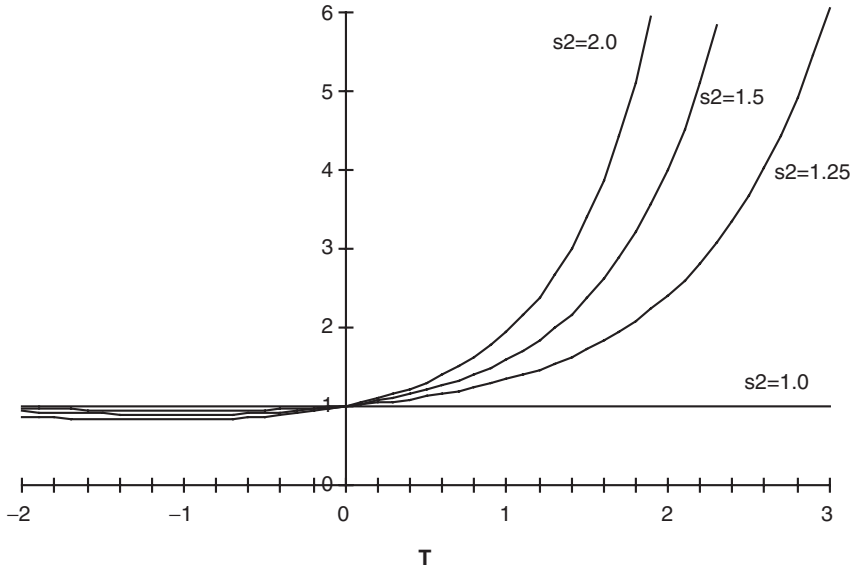
Now let us simplify in another way by considering what happens when we have two groups of equal means but unequal standard deviations, that is, $a_1 = a_2 = a$, and $s_1 < s_2$.

For the case of unequal standard deviations and equal means, when the hurdle is a very low one relative to the group means, differences in the proportions of each of the two groups that gain acceptance can slightly advantage the group with the lower standard deviation. However, as the magnitude of the hurdle increases, even modest differences in standard deviation can make dramatic differences in the proportion of each group that passes the test. Indeed, when the hurdle is very high, that is, far enough out "on the tails" of each distribution, we will find that representation of the group with low standard deviation is minuscule or zero, despite the fact that its mean is identical to that of the other group. We illustrate this point graphically in Figure 4, showing the success ratio for $a = 0$ and $s_1 = 1$, for various values of s_2.

Let us return to the real data we previously looked at for choosing members of a campus honor society (see Table 1). If we compare, say, unaffiliated students (which we will now call Group 1) with biological science majors (Group 2), we find that they have similar means but different standard deviations. If the threshold is, say, 3.2, then the success ratio is 1.22. If the threshold is as high as 3.80, the success ratio is 2.06, that is, Group 2 members *are more than twice as likely* to meet the criterion. This empirical

FIGURE 4

Success Ratio as a Function of T for Fixed $s_1 = 1$, $a_1 = a_2 = 0$, and Selected Values of s_2



example, using real data, shows that differences in success rates of a nontrivial magnitude can arise from what appear to be relatively minor differences in *standard deviations* between groups.

Different Means and Different Standard Deviations

We now investigate the tradeoff when both means and standard deviations differ between the groups. Assume without loss of generality that $a_1 < a_2$. Denoting by $S_1(T)$ and $S_2(T)$, respectively, the proportions of Groups 1 and 2 who succeed, we first determine values of T for which an equal proportion succeed, that is, for which $S_1(T) = S_2(T)$. Because the distribution of each group is normal, the standard scores of T must be the same for each distribution, that is:

$$\frac{T - a_1}{s_1} = \frac{T - a_2}{s_2}.$$

Solving for T, we obtain:

$$\bar{T} = \frac{a_2 s_1 - a_1 s_2}{s_1 - s_2} = \frac{a_2 R - a_1}{R - 1},$$

where $R = s_1/s_2$, and we have denoted the unique threshold that yields equal-success proportions by $\bar{T}$. This *critical threshold*, $\bar{T}$, depends only on the locations of the group means and the ratio of the standard deviations.

If $s_1 > s_2$ and a threshold, T, is greater than the critical value, $\bar{T}$, then Group 1 (the group with the lower mean but greater standard deviation) is advantaged; if T is below $\bar{T}$, Group 2 is advantaged. When $s_1 > s_2$, it is easy to check (using the fact that $a_1 < a_2$) that $\bar{T} > a_2$. In fact, if s_1 is only slightly greater than s_2, then $\bar{T}$ is very large and Group 2 is advantaged for most thresholds. When s_1 greatly exceeds s_2, however, the value of $\bar{T}$ is close to a_2. Conversely, if $s_1 < s_2$, Group 2 is advantaged when the threshold exceeds $\bar{T}$ but not when it is lower than $\bar{T}$. In this case $\bar{T}$ is less than a_1, approaching it when s_1 is very small in comparison to s_2.

Finally, some algebra shows that the common proportion who succeed in each group when the threshold is set at the critical value, $\bar{T}$, is given by:

$$S_1(\bar{T}) = S_2(\bar{T}) = 1 - \Phi\left(\frac{a_2 - a_1}{s_1 - s_2}\right),$$

where Φ is the standard cumulative normal distribution function.

It follows that if $s_1 > s_2$, the proportion succeeding at the critical threshold will decrease as the group means become more separated but will increase as the group standard deviations become more disparate. In this case, that is, when Group 1 has the larger standard deviation, the proportion succeeding will be less than 0.5 (because $\bar{T}$ is to the right of a_2) and typically much less. On the other hand, if Group 1 has the smaller standard deviation, the reverse relationships will hold, and the proportion succeeding will exceed 0.5. It is the former case—in which a relatively small proportion succeed—that is most relevant to the applications that we consider in this article.

We show in Figure 5 probability densities and probabilities of success for groups with disparate means and standard deviations.[10]

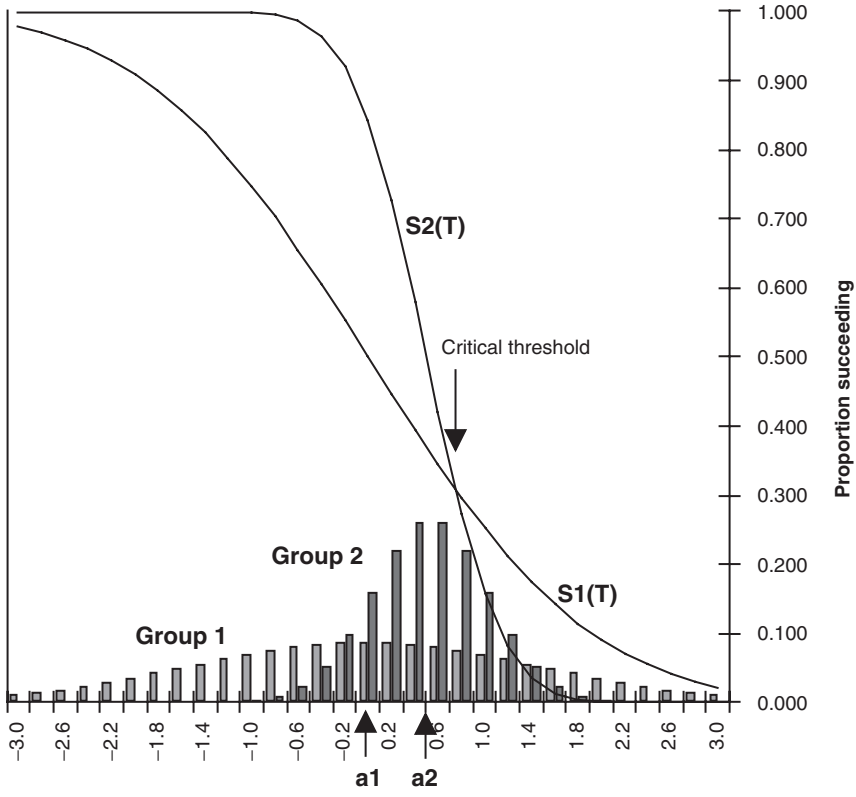
Lottery-Based Admissions with a Variable Threshold Versus Race-Norming as Tools for Affirmative Action

The low representation of historically disadvantaged groups such as African Americans in colleges and universities, as well as in certain occupations and in positions of authority more generally, has inspired great public concern in the United States. A variety of devices, falling under the general rubric of "affirmative action," have been adopted to increase minority

[10]In this context, we might note that Jerome Bruner, commenting on an essay by Marshall Smith, Professor of Education at Stanford University, noted, only partly jokingly, that although American students do not do well on international tests of knowledge and skills, "where we undoubtedly lead the world is in variability. American standard deviations on all the tests we take are just about at the top. . . . America seems to have a gift for fostering . . . inequality" (Bruner, 2003:51).

FIGURE 5

Success Functions and Critical Threshold for Groups with Disparate Means ($a_1 = 0$, $a_2 = 0.5$) and Disparate Standard Deviations ($s_1 = 1.5$, $s_2 = 0.5$)



success in admissions and in the employment arena.[11] These devices range from simple attempts to publicize opportunities better among the members of the minority community, to giving preference to the minority candidate when candidates are otherwise regarded as equally qualified, to applying different standards for admissions (or hiring or promotion) to members of different groups (so-called *race norming*), to admitting applicants (or allocating jobs) on the basis of racial quotas. The further the deviation from what is seen as ''hiring the best qualified'' (defined, of course, perhaps quite

[11]Gender differences in test scores are also important. There is considerable evidence that the key racial differences in test scores such as SATs tend to be in the means, while gender differences often are found in standard deviations as well, sometimes even where there is no real difference in means (see, e.g., Bridgeman and Lewis, 1996; Hedges and Nowell, 1999).

misleadingly, in terms of test scores/credentials), the more controversial are proposed techniques of affirmative action.[12]

In this section, we focus on the comparison of two different schemes for minority preference: *race-norming* and an important scheme that we have labeled *lottery-based rules with minimum thresholds of acceptance*. Our concern will be to elucidate the statistical relationships that will determine the effects that lottery-based minimum threshold schemes will have on the relative success rates of minority and nonminority applicants when compared to race-normed schemes, on the one hand, and to test-score-based admission schemes that simply admit those who score best, on the other.

Race-norming sets the thresholds of acceptance for each group differently; in its most extreme form, these thresholds are chosen so that the same proportion of each group will score success (proportional quotas). Race-norming of standards is found in many places; it is especially common in educational admission. For example, in one ''magnet'' high school in San Francisco, Lowell High, because test scores of Chinese-American students were, on average, so much higher than most of those in other groups, in order to assure more racial diversity in its student body, this high school required higher test scores for admitting students of Chinese-American descent than for any other students, with the next highest thresholds being used for students of Asian-American but not Chinese-American background and for non-Hispanic whites. The lowest thresholds were for African Americans and for students of Hispanic/Latino heritage. The difference between the cut-off value used for Chinese Americans and that used for African Americans was quite extreme.[13] However, at Lowell High, as a result of a 1994 lawsuit filed by Chinese-American parents, straight race-norming of applications was initially replaced by a combination of straight test scores and GPA (for 80 percent of the students) and eight special weighting criteria for 20 percent of the class-slots criteria that operate to favor Hispanic and African-American students (Luis Fraga, personal communication, June 17, 1998), and more recently by a complex formula that is referred to as ''diversity enhancing,'' where diversity is defined by student and family background characteristics that include language use in the home, whether the mother has graduated from high school, whether the student is in

[12]In the employment arena, various forms of minority preferences are justified as (1) necessary equitable remedies for past histories of discrimination, especially for firms that have been legally found guilty of such discrimination; (2) given the argued persistence of sub rosa and institutional discrimination, needed means of compensation for continuing employment practices based on tests that supposedly do not usefully measure likely job performance and are biased against minorities; and/or (3) an appropriate way to reflect our national diversity.

[13]Chinese-American students had to score 62; non-Hispanic whites and non-Chinese Asians had to score 58; African Americans and Hispanics needed only score 53 (Petersen, 1997:62). The form of race-norming that was used did not actually achieve proportionality in the school since students of Chinese-American and Asian-American descent were still overrepresented relative to the applicant pool. Moreover, there was a further complication in the form of a court-mandated requirement that no racial group could constitute more than 40 percent of the school's admittances.

public housing or eligible for the free lunch program, and so forth (Fletcher, 2002).

Lottery-based rules set a threshold for acceptance (presumably at or above the minimum level of competence needed for satisfactory performance) and use a lottery to choose among applicants who score above that threshold regardless or race or ethnicity or gender and regardless of differences in their scores. As noted earlier, such rules have recently been proposed (Cohen, 1994; Guinier, 1996) as a way to mediate between the seemingly irreconcilable positions of those who advocate race-normed standards and those who advocate so-called merit (i.e., test-score-based) decision making to be applied in a color-blind fashion. Indeed, Guinier (1996) specifically suggested such rules might be used to resolve the conflict over admissions criteria at Lowell High School.

Outside the United States, lotteries (with minimum thresholds) have been used for pupil assignments so as to reduce SES differences in admittance that would arise from test-based procedures. For example, in the Netherlands, there is "a weighted system under which all applicants who pass a moderate grade threshold in specified subjects participate in the lottery, but those with higher grades have more entry forms put in for them" (Heidenheimer, Heclo, and Adams, 1990:50, internal cites omitted). Similarly, Germany in the 1980s began to determine some of its medical school admission decisions in this way (Heidenheimer, Heclo, and Adams, 1990:50). Lottery-based schemes with minimum threshold may achieve greater importance as a tool for affirmative action if race-norming suffers further popular disapproval via referenda (such as Proposition 209 in California in 1996 or a similar referendum in 1998 in the State of Washington).[14]

Without great loss of generality, we shall focus on pure lottery rules with minimum thresholds rather than rules that mix this method with others.[15] Suppose a (relatively low) threshold, T_1, is used to establish a pool of acceptable applicants, from whom a smaller subset is chosen by lottery for admission. Since the lottery has no effect on the (expected value of the) proportions of the acceptable pool represented by each recognized subgroup, the relative proportions of subgroups accepted for admission depend entirely

[14]In its 2003 decisions about affirmative action admissions policies at the University of Michigan (*Grutter v. Bollinger*, 539 U. S. 982, dealing with law school admissions to the University of Michigan, and *Gratz v. Bollinger*, 538 U.S. 959), while the Supreme Court rejected the points system used for undergraduate admissions, in holding the University of Michigan Law School admission procedures to be constitutional, the Supreme Court majority endorsed the view that diversity was a factor that could be taken into account. However, exactly how diversity was to be defined and how much it could be allowed to matter were matters left still open. Most importantly, while diversity goals are permitted in higher education if they are narrowly tailored, they are not constitutionally required. Thus, we anticipate that partially lottery-based admission procedures of the sort we discuss will remain of interest to reformers who wish to achieve diversity but without mechanisms that explicitly take race or ethnicity or gender into account.

[15]We can analyze the effect of such mixed rules by treating them as weighted combinations of the pure types.

on the threshold, T_1. Under this plan a less privileged group would be admitted in greater proportions than under a race-blind procedure defined by a single, higher threshold, T_2, simply because the success ratio is an increasing function of T.[16] In fact, the success ratio—from the point of view of the less privileged of two subgroups—would be improved by the factor:

$$SR(T_2)/SR(T_1),$$

that is, the quotient of the success ratio for the two groups at the two thresholds.

A pure race-normed procedure, on the other hand, produces, by design, admissions in direct proportion to group membership in the population. It thus achieves a more proportional outcome vis-à-vis descriptive representation than lottery-based admissions but with the overt discrimination inherent in group-specific thresholds. Thus, with an appropriate threshold chosen, lottery-based rules produce outcomes that are intermediate in descriptive representation between pure race-normed or gender-normed standards for acceptance, on the one hand, and the decision to simply select those with the highest score regardless of race or gender, on the other.

However, unless thresholds are set very low, lottery-based schemes may be limited in how close they come to proportionality. In such a situation, with a moderately high threshold (e.g., one high enough to screen out those who have no realistic chance at success) the vast bulk of those who score above this threshold may still be members of the more advantaged group. Even if the subgroup whose members have highest mean hurdle-overcoming abilities is small in size it is possible that a lowered threshold may enable a very large proportion of its members to be eligible for the lottery.

Suppose, for example, the population consists of two groups, with the second group having the higher mean, and that, given the number of spaces to be filled, we will just fill all the spaces if we select 10 percent of the combined population. The same number of admittances would be accomplished by setting the race-normed thresholds, T_1 and T_2, as the 90th percentiles of the respective groups and admitting everyone above those percentiles in each group, respectively. On the other hand, by setting the lottery-based threshold as T_1, that is, by using the same minimum threshold of admission for the entire population as was used to screen members of the less advantaged group, 10 percent of the first group make the first cut, while a larger fraction, let us say 40 percent, of Group 2 make this cut. Now, of course, we will select from this set a random proportion equal to 10 percent of the population of the initial application pool. Because of the lottery, the same relative fractions make the final selection, that is, the success ratio is four times the size ratio between the two groups,

[16]We assume here that the total number of admissions is fixed.

that is:

$$SR = \frac{n_2 p_2}{n_1 p_1},$$

where n_i is the size of group i and p_i is the proportion of group i making the first cut.

Thus, the relative proportion selected from the two groups differs from size proportionality by the factor, p_2/p_1, which is at least 1 and will tend to be much higher if the group densities overlap little. In other words, if the more advantaged group has a much higher proportion of its members who can pass the lower threshold, T_1, than is true for the less advantaged group, lowering the threshold to reflect the minimal scores sufficient for acceptable candidates to reflect those used to determine admissions for the less advantaged group and then using a lottery may still give rise to greatly discrepant acceptance rates among members of the less advantaged and the more advantaged groups. Moreover, the mean score of those admitted will fall compared to the race-normed plan since now the members of Group 2 who will be admitted have on average lower scores than before (we are selecting from the pool of people above T_1 rather than the pool of those above T_2).[17] Note also the important result that the relative size of groups is immaterial in these conclusions (see example above).

When a single "group-blind" (i.e., color blind, gender blind, etc.) threshold of acceptability is used for both groups, such a standard necessarily lies between T_1 and T_2, and the success ratio may be far worse from the perspective of Group 1 than under the lottery plan. But, in the example above, relative to a "group-blind" system, a lottery-based threshold system (restricted to 10 percent of the total population) does not, in fact, change the representation of the least represented group by that much ($SR = 2.1$ vs. $SR = 2.3$).

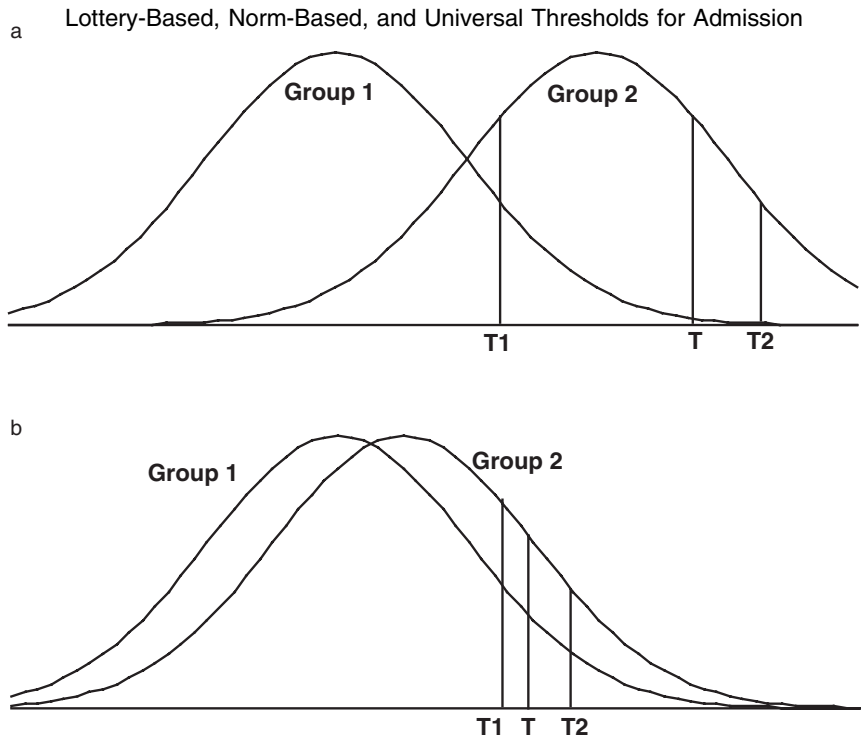
Whether the results of a lottery-based selection more resembles in descriptive representation that of a group-specific or a universal threshold selection procedure depends on the relative positions of the two distributions, as is illustrated in Figure 6. In Figure 6a, the universal threshold would be closer to T_2 than to T_1; in Figure 6b, the reverse would be true.

Figure 6 has important practical implications: the advocacy in the abstract of a lottery-based minimum threshold by spokespersons for "minority" representation may turn out not to be very efficacious in increasing some particular group's proportion of the admittance pool in settings where substantial differences in group means obtain—unless the minimum threshold of acceptance is set so low that it may not be seen as sufficiently meritocratic.

For example, returning to the actual GPA data we discussed previously, suppose we plan to select 10 percent of an overall population consisting of

[17]However, lottery methods may yield lower overall standard deviations than race-norming.

FIGURE 6



NOTES: a. Disparate groups. Norm-based thresholds T_1 and T_2 admit 10 percent of the respective groups. If T_1 is used for a lottery-based threshold, the success ratio is about 8 to 1. To achieve the same overall admissions, a universal threshold, T, must be placed nearer to T_2.
b. Similar groups. Norm-based thresholds T_1 and T_2 admit 10 percent of the respective groups. If T_1 is used for a lottery-based threshold, the success ratio is about 2 to 1. To achieve the same overall admissions, a universal threshold, T, must be placed nearer to T_1.

engineering (Group 1) and humanities students (Group 2), assuming for simplicity an equal number in each. Under a lottery-based selection, T_1 (= 3.22) and T_2 (= 3.51) are chosen as the 90th percentiles of the respective groups. Then $p_1 = 0.10$ and $p_2 = 0.21$ (i.e., 21 percent of the humanities students pass T_1), so that $SR = 2.1$. By contrast, under a group-normed system, $SR = 1.0$, while under an individualistic (i.e., group-blind) system, $SR = 2.3$. Here a lottery system (with minimum threshold) does not substantially raise the proportions of members of the disadvantaged group who quality.

According to the Educational Testing Service webpage, in 1997, college-bound non-Hispanic whites had mean SAT verbal scores of 526 ($SD = 101$) and mean SAT math scores of 526 ($SD = 103$). College-bound African Americans had mean SAT verbal scores of 434 ($SD = 101$) and mean SAT

math scores of 423 ($SD = 97$). Given the actual composition of SAT test-takers by race, if we wished to select the top 10 percent of scorers on the verbal part of the test, using a universal threshold, the proportion of whites who will surmount that threshold is 5.5 times the proportion of African Americans with scores that high. For the math SAT scores, because of the greater separation between means and the smaller *SD*s, the result for a top 10 percent selection is even more exaggerated; the proportion of whites who fall into that category is 8.5 times higher than the proportion of African Americans who fall into that category.

Using a 10 percent cutoff, as noted above, the group-blind procedure had a white/black success ratio of 5.5 for the SAT verbal and 8.5 for the SAT math. Of course, the group-normed procedure must have a success ratio of 1.0. What about the lottery-based procedure with minimum threshold set at the SAT score needed for the norming of African-American scores to achieve a 10 percent success rate (here 563 verbal and 547 math)? For this lottery-based selection with minimum threshold, for the SAT verbal scores, 35.7 percent of the white students pass the 563 threshold while only 10 percent of African-American students pass—yielding a white/black success ratio of 3.6. For the SAT math scores, the corresponding lottery-based threshold yields a success ratio of 5.5. Here, a lottery-based system with a realistic minimum threshold will result in only a minuscule rate of minority acceptance compared to that of whites.

Another important issue has to do with the extent of reduction in mean test scores among those applicants chosen by the procedure as we shift from a pure test-score-based to a pure proportional-quota-based system (cf. Berger, Wang, and Monahan, 1998).

In our previous GPA example, if we set the universal threshold at $T = 3.39$, the mean score for the student who exceeds this threshold is 3.68. If a group-normed procedure is used with thresholds set to give rise to equal proportions of each group (on average) being selected (here with thresholds at 3.22 and 3.51, respectively), the mean score among successful applicants is 3.65, quite close to that of the group-blind system. On the other hand, for a lottery-based system (with threshold set at $T_1 = 3.22$, the lower of the two group-normed thresholds), the expected mean score for the selected students drops to 3.54. Thus, the lottery-based system yields the lowest mean test score among those who are chosen.

The latter result goes through even when the groups we are comparing are very different in means. If we look at SAT scores broken down by race, we begin with rather substantial mean differences, as we saw earlier. Again, consider an elite institution that uses a rather high SAT threshold to screen applications in a group-blind (universal) fashion, taking the top 10 percent. For SAT verbal scores, this universal threshold is 623; for SAT math scores it is 621. The mean score for the students who exceed this universal SAT verbal threshold will be 674, and the same mean score obtains for the students who exceed the universal SAT math threshold.

If a group-normed procedure is used, we get non-Hispanic white SAT thresholds at 655 and 658, for SAT verbal scores and math scores, respectively; and comparable African-American thresholds at 563 and 547. For this group-normed procedure, the mean SAT scores of successful applicants are 657 verbal and 650 math. This is a drop of roughly 20 points in each, on average, as compared to the mean scores for admittance under a universal threshold. It is not as large as we might have thought because the raising of the threshold for non-Hispanic whites compensates in part for the lowering of the threshold for African Americans.

For a lottery-based system (with minimum threshold set at 563 for SAT verbal scores and 547 for SAT math scores, the scores used previously to determine African-American admittances under a group-normed scheme), the expected mean scores for the selected students drop to 627 for verbal scores and to 617 for math scores. Again, the lottery-based scheme has the most severe consequences for the mean test scores of those who are chosen. Moreover, the drops shown in these SAT examples are quite substantial—a drop of around 50 points in each, on average, as compared to the mean scores for admittance under a universal threshold.

Thus, lottery-based schemes (with minimum thresholds) must be scrutinized closely for their likely effects. On the one hand, they may not improve group representation in descriptive terms much over pure test-based rules; and they may lead to a severe drop in mean test scores of the successful applicants as compared to either test-based rules or group-norming. Unlike a group-normed system, a lottery system adds to the pool of applicants eligible for (at risk for) selection not only those in the less advantaged group who meet the lower threshold—just as in a group-normed system—but also the larger number of persons in the more advantaged group that meets this threshold. The effect is that a lottery system tends to lower the mean score of successful applicants far more than a group-normed system. We do not believe that this feature of lottery systems is well understood by those who might advocate them in preference to race-norming.

Discussion

The models we have provided help us understand why the differences among subpopulations on test scores or on performance variables will have quite different consequences depending on the threshold that is used as the cutoff for admissions/hiring or for promotion/honor society membership, and so forth. When we turn to the recently proposed affirmative action remedy of a lottery-based rule with minimum threshold, our analysis provides the necessary guidelines to evaluate the likely impact of such a rule relative to either pure highest-test-score admission rules or race-norming. Our SAT example shows that, when groups differ significantly in test scores,

such lottery rules may not yield results that are particularly close to proportionality unless the threshold for acceptance is set unreasonably low, and that, even with a minimum threshold, use of a lottery can have a significant effect on the mean score of those who are accepted. Thus, while a lottery-based method avoids explicit race-norming, it can impose other costs.

Although we have focused here only on three basic types of procedures; universal, race-normed, and lottery rules with minimum thresholds, the approach we offer is considerably more general.[18] Of course, the analyses we have presented would need to be modified to take into account the complexities of various real-world allocational rules. For example, multiple tests might be used. Or test scores might be used for some openings, with others being filled through lottery-like procedures, while other spaces are filled in ways that are explicitly attentive to diversity concerns. Or individuals might be given composite scores that combine test information with more subjective judgments (e.g., based on personal interviews). Nonetheless, we believe that, in the context of affirmative action, it is clear that the basic logic of statistical thresholds that we have exposited in this article will have important uses for policy analysis and the analysis of institutional design when we are evaluating allocational schemes that make *any* use of tests on which groups may differ in their members' means and/or standard deviations. Moreover our results are highly relevant for legal analyses of affirmative action questions, especially as they have to do with whether some particular affirmative action scheme is "narrowly tailored" to achieve its goals.[19]

[18]Threshold-related models have been generated in a number of different disciplines—often in ignorance of previous work done in other subfields. Models mathematically similar to those we use have been applied in a number of different substantive contexts, ranging from signal detection in cognitive psychology (Luce, 1965); to modeling the impact of changes in juror thresholds for deciding guilt beyond a reasonable doubt on Type I and Type II error rates (Nagel and Neef, 1975; Grofman, 1981); to modeling collective action (e.g., riot participation) as a sequential process (Granovetter, 1978; Granovetter and Soong, 1983; Macy, 1991; Lohmann, 1994, 1997); to modeling the impact of changes in voter registration laws (interpreted as barriers to participation) on the voter turnout of different socioeconomic groups (Brians and Grofman, 1999; cf. Erikson, 1981; Cox and Munger, 1990; Fort, 1995). Space constraints have prevented us from exploring the similarities and contrasts between these models and our own work on dichotomous-choice threshold models.

[19]In this context we should note that the University of Michigan Law School reviewed the option of using a lottery-based rule with minimum threshold and rejected it in favor of what the Supreme Court majority in *Grutter* labeled more "holistic" approaches. Professor Richard Lempert (University of Michigan Law School) testified at the trial challenging the law school's admission criteria that a system in which the law school would lower its admissions standards, establish a numerical cutoff for "qualified" applicants, and then select randomly among those applicants, "would admit greater number of minority students, but would not yield meaningful racial and ethnic diversity" (cited in *Grutter v. Bollinger et al.*, 2002 Fed. App. 0170P (6th Cir. en banc, Dec. 6, 2001)). In the Supeme Court decision in this case, Justice O'Connor, writing for the majority, indicated that she believed the law school plan had "adequately considered the available alternatives," including a "lottery system."

REFERENCES

Alt, James. 1994. "The Impact of the Voting Rights Act on Black and White Voter Registration in the South." In Chandler Davidson and Bernard Grofman, eds., *Quiet Revolution in the South*. Princeton, NJ: Princeton University Press.

Barlow, Richard, and Frank Proschan. 1965. *Mathematical Theory of Reliability*. New York: J. Wiley & Sons.

Berger, Paul D., Chen Wang, and James P. Monahan. 1998. "Quantifying a Statistical Aspect of Segmented Selection/Quota Systems." *American Statistician* 52:228–32.

Brians, Craig, and Bernard Grofman. 1999. "When Registration Barriers Fall, Who Votes? An Empirical Test of a Rational Choice Model." *Public Choice.* 99:161–76.

Bridgeman, B., and C. Lewis. 1996. "Gender Differences in College Mathematics Grades and SAT-M Scores—A Reanalysis of Wainer and Steinberg." *Journal of Educational Measurement* 33(3):257–70.

Bruner, Jerome. 2003. "Comment on 'Education Reform: A Report Card'." *Bulletin of the American Academy of Arts and Sciences* 56(2):38–52.

Cohen, Richard. 1994. "'Bounding' of Test Scores as a Merit-Based Remedy for Employment Discrimination." Remarks at Joyce Foundation Conference on The Civil Rights Act of 1964 in Perspective. Washington, DC: National Judicial Center.

Cox, Gary, and Michael Munger. 1990. *Putting Last Things Last: A Sequential Barriers Model of Turnout and Voter Roll-Off.* Unpublished manuscript. Department of Political Science, University of North Carolina.

Crow, James F. 2002. "Unequal by Nature: A Geneticist's Perspective on Human Differences." *Daedalus* Winter:81–88.

Erikson, Robert. 1981. "Why Do People Vote? Because They Are Registered." *American Politics Quarterly* 9(3):259–76.

Fletcher, Michael A. 2002. "Desegregation, But Not by Race: San Francisco Turns to Socioeconomic Factors to Balance Schools." *Washington Post National Weekly Edition* March 25–31:20.

Fort, Rodney. 1995. "A Recursive Treatment of the Hurdles to Voting." *Public Choice* 85 (1–2):45–69.

Granovetter, Mark. 1978. "Threshold Models of Collective Behavior." *American Journal of Sociology* 83:1420–33.

Granovetter, Mark, and Roland Soong. 1983. "Threshold Models of Diffusion and Collective Behavior." *Journal of Mathematical Sociology* 9:165–79.

Grofman, Bernard. 1981. "Mathematical Models of Juror and Jury Decision-Making: The State of the Art." Pp. 305–51 in Bruce D. Sales, ed., *Perspectives in Law and Psychology, Volume II: The Trial Process*. New York: Plenum.

Gross, Alan, and Virginia Clark. 1975. *Survival Distributions: Reliability Applications in the Biomedical Sciences*. New York: John Wiley & Sons.

Guinier, Lani. 1996. "The Need for a National Conversation on Race." Remarks at the E Pluribus Unum Conference, Program in Chicano Studies, Stanford University. Stanford, CA.

Guinier, Lani, Michelle Fine, and Jane Balin. 1997. *Becoming a Gentleman: Women, Law School and Institutional Change*. Boston, MA: Beacon Press.

Hedges, Larry V., and Amy Nowell. 1999. "Changes in the Black-White Gap in Achievement Test Scores." *Sociology of Education* 72:111–35.

Heidenheimer, Arnold J., Hugh Heclo, and Carolyn Teich Adams. 1990. *Comparative Public Policy*, 3rd ed. New York: St. Martins Press.

Jencks, Christopher, and Meredith Phillips, eds. 1998. *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.

Kadane, Joseph B., and Caroline Mitchell. 2000. "Statistics as Legal Proof in Employment Discrimination Cases." In Bernard Grofman, ed., *Legacies of the Civil Rights Act of 1964 in Perspective*. Charlottesville, VA: University Press of Virginia.

Kahlenberg, Richard D. 1996. *The Remedy: Class, Race, and Affirmative Action*. New York: Basic Books.

Kalbfleisch, John D., and Ross L. Prentice. 1980. *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons.

Kane, Thomas J. 1998. "Racial and Ethnic Preferences in College Admissions." Pp. 431–56 in Christopher Jencks and Meredith Phillips, eds., *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.

Lohmann, Suzanne. 1994. "Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989–1991." *World Politics* 47:42–101.

———. 1997. *Dynamics of Information Cascades: Experimental Evidence*. Unpublished manuscript. Department of Political Science, UCLA.

Luce, R. Duncan, ed. 1965. *Handbook of Mathematical Psychology*. New York: Wiley.

Macy, Michael W. 1991. "Chains of Cooperation: Threshold Effects in Collective Action." *American Sociological Review* 55:730–47.

Meier, Paul, Jerome Sacks, and Sandy L. Zabell. 1984. "What Happened in *Hazelwood*: Statistics, Employment Discrimination and the 80% Rule." *American Bar Foundation Journal* 1:139–86.

Nagel, Stuart S., and Marian Neef. 1975. "Deductive Modeling to Determine Optimum Jury Size and Fraction Required to Convict." *Washington University Law Review* 97:933–78.

Petersen, William. 1997. *Ethnicity Counts*. New Brunswick, NJ: Transaction Publishers.

Poulos, John Allen. 1995. *A Mathematician Reads the Newspaper*. New York: Basic Books.

Rae, Douglas et al. 1981. *Equalities*. Cambridge, MA: Harvard University Press.

Spence, A. Michael. 1974. *Market Signaling: Informational Transfer in Hiring and Related Screening Processes*. Princeton: Princeton University Press.

Zwick, Rebecca. 2002. *Fair Game? The Use of Standardized Admissions Tests in Higher Education*. New York/London. Routledge Falmer.