Psychology, Public Policy, and Law
2000, Vol. 6, No. 1, 233–252

Copyright 2000 by the American Psychological Association, Inc.
1076-8971/00/$5.00 DOI: 10.1037//1076-8971.6.1.233

SO NEAR AND YET SO FAR
Lingering Questions About the Use of Measures of General Intelligence for College Admission and Employment Screening

Stephen J. Ceci
Cornell University

The evidentiary bases for the various positions taken by the authors of articles in the March 2000 special theme issue of *Psychology, Public Policy, and Law,* are examined. Despite some substantive disagreements among the authors, a substrate of evidence is accepted (or, at the very least, goes unchallenged) by all authors. Although the substantial progress made by psychometric researchers is acknowledged, there remains a critical need to probe ever more deeply into the nature of intelligence and the meaning of correlations between ability tests and performance at school and at work. The author concludes by arguing that the concept of general intelligence may have considerable predictive usefulness whenever the situation calls for making a limited number of choices among many applicants but that there are lingering conceptual limitations about its meaning.

As an avid reader and researcher of intelligence (what does it mean, what does it predict, and why does it predict), I accepted Wendy Williams's invitation to discuss the 18 articles in this special theme issue with only a modicum of hope. Perhaps I would learn something new, I told myself, but experience as an author, commentator, and editor in the field of intelligence led me to anticipate a low yield of new knowledge from reading the articles in this issue. I was wrong. *Very* wrong. I was surprised to discover that there is indeed much that is new in this special issue. There is much that may challenge our cherished beliefs about the existence of individual and group differences and their import for selection into schools and jobs. Because of the high level of scholarship exhibited by the authors of every article in this issue, it is imperative that their arguments be taken seriously. Readers who are familiar with the field of intelligence testing will find these articles fascinating; readers who are unfamiliar with the subject will find these articles provocative, even upsetting at times, but always interesting.

In what follows, I have organized my reactions to the articles in terms of the categories listed by Wendy Williams in her introduction to the issue (Williams, 2000). Admittedly, these are overlapping and somewhat arbitrary. To adumbrate my conclusions, I believe that there currently exists substantial agreement among researchers about a number of findings, although the interpretation given to some of them is still open to debate.

Correspondence concerning this article should be addressed to Stephen J. Ceci, Department of Human Development, NG25 Martha Van Rensselaer Hall, Cornell University, Ithaca, New York 14853. Electronic mail may be sent to sjc9@cornell.edu.

Part I: Intelligence Testing in Historical Perspective

Articles by Sheldon White (2000) and Alvin Calvin (2000) set the stage for this special issue by putting modern controversies about nature–nurture into historical and contextual perspective. Williams deserves great credit for asking these scholars to contribute articles to this special issue because neither has made their reputation in this area of scholarship, yet, as is plainly evident, both have important things to contribute. Calvin's first-hand report of the development of the SAT is extremely valuable. Although I regard myself as knowledgeable in this area, I was not aware of the seminal role played by Henry Chauncey in the popularization of the Scholastic Assessment Test (SAT) nor of the impetus provided by the then-president of Harvard, James Bryant Conant. Reading Calvin's essay is like peeping into an earlier epoch, when values that we now regard as so thoroughly mainstream as to be hackneyed were not yet seen as self-evident.

In the final pages of his article, Calvin proposes a lottery plan similar to the one used in the Netherlands to decide which of the right-tail (i.e., highly qualified) applicants get admitted to elite colleges and universities. Cognizant of the objections to such a proposal, Calvin nevertheless puts it forward to spur debate in the hope that a nonadversarial discussion can lead to consensus and resolution of the seemingly intractable factions that have formed in response to affirmative action initiatives. Although his examples are the stuff that wins debates (specifically, his sense of unfairness that his own Asian American grandson would need to score 200 points higher on the SAT than his other grandson who happens to be a Latino, to gain entry to the University of California at Berkeley), the Netherlands lottery is no panacea: Anytime we turn continuous data into cutoff scores we lose predictiveness, at least if the regression is homogeneous across the entire spectrum. Still, against the backdrop of a system in which "disparate impact" per se (i.e., the mere existence of racial, ethnic, or gender differences in the outcome of admissions or hiring decisions) is taken as evidence that the test itself is invalid, Calvin's plan is superior to race norming, and I believe that it avoids some of the sticky problems that inhere in the more race-conscious proposal outlined by Perloff and Bryant in the summation to their provocative article (Perloff & Bryant, 2000).

In the second article of this section, Sheldon White (2000) provides a rich and revealing description of the origins of the early testing movement. Of particular importance is his demonstration that the first U.S. test developers were animated by the assumption that mental subnormality was but one part of a larger constellation of degenerative proclivities that included criminality proneness, religious fanaticism, vagrancy, sexual excesses, and poor physical health. White makes the point that the origins of U.S. and European thinking about intelligence were grounded in pseudoevolutionary ideas about the ideal endpoint of intellectual development, the summum bonum of mental evolution as he puts it.

Toward the end of his article, White lists eight pithy propositions that are, for the most part, uncontroversial. The first of two exceptions to this conclusion is his assertion that even if races, classes, or cultural groups possess different mean IQs, this is "neither terribly surprising nor critical in an educational or social sense" (p. 39). White's friend and colleague, Dick Herrnstein, argued in *The Bell Curve* that if two groups' mean IQ differed by a full standard deviation, then this could lead to significant and surprising social outcomes. For example, if a state employment office requested the applications of only those individuals whose IQs were above 115 (i.e., 1 *SD* above the White mean but 2 *SD*s above the Black mean), then one might as well post a sign saying, "Blacks need not apply" because the realities of the bell curve will mean that relatively few Black applicants will be in the pool (roughly 5 of 6 applicants will be White because 115 represents 2 *SD*s above the Black mean, so only approximately 3% of Black scores will fall above this cutoff score, whereas approximately 16% of White scores will do so). This fact surprises my students, and, I suspect, it is surprising to policy makers and legislators as well.

White's other controversial claim is that we do not know how to make IQ tests culture fair. As Reynolds' (2000) and Halpern's (2000) articles demonstrate, reasonable scholars can and do disagree about the meaning of culture fair. To some individuals, it is enough that a predictive equation for Whites predicts equally well for Blacks (or in the case of Blacks actually overpredicts); to others, such predictiveness does not speak to the issue of whether racial groups have equal access to test-related experiences and materials. So, in a way, White is right about the futility of making tests culture free, but some would say that this still leaves room for them to be culturally fair.

Neither of these points diminishes White's excellent contribution to this special issue. His sense of history is impeccable, and his reasoning reminds us that the most incendiary aspects of the debate about intelligence have been ignited by the choice of words whose meanings predate the inception of IQ tests. For this, we can thank Sheldon White.

Part II: Assessing the Value of Affirmative Action (AA)

The articles in this section are among the most provocative in the entire issue, and they are sure to generate discussion among policy makers and representation from the media. Douglas K. Detterman (2000) sets it up with his sharp and novel argument that AA has not achieved one of its primary goals, to equalize graduation rates. To date, no one has tackled AA on this basis, and it deserves to be examined closely. Linda Wightman (2000) takes a different tack resulting in a position that is nevertheless compatible with the one advocated by Detterman, namely, that high educational goals and diversity can coexist with the use of standardized test scores. Halpern wades in with a very different view of the meaning of aptitude and fairness; whereas, articles by Howard Everson (2000) and Robert Perloff and Fred Bryant (2000) add to the mix by endorsing programs that, in the former case, are consistent with current policies, and in the latter case, go beyond them by proposing a more race-conscious plan that starts with a goal of the number of minority students or workers that an institution wishes to have and then uses a set of weights to select the top minority and majority applicants. Finally, Scullin, et al. (2000) provide some suggestive evidence that AA is working fairly pervasively and that the group most underremunerated after controlling for education and IQ are White women.

Detterman does a great service by bringing the argument back to data, and he provides several pieces of evidence for his claim that AA has not worked. It is fascinating to pit Detterman's article against the two others in this section, Wightman's and Halpern's, as well as against an earlier article by Wightman (1997) that appeared in the *New York Law Review.* Together, these three articles tell a complex but extremely important story.

Let's start with three findings from Wightman's 1997 article that help frame the discussion of Detterman's proposal for open admissions. First, Wightman shows that AA is taken seriously by law schools because minority applicants are admitted at the same rate as Whites, despite substantial differences in test scores. Second, through data she shows that the use of any alternative to AA, such as socioeconomic status (SES), academic major, or status of undergraduate institution, would result in substantial minority disparities in admission. Third, Wightman shows that there are differences in outcome for the various groups: Although there is no statistically significant difference in passing the bar within each racial–ethnic group, there are substantial differences between racial–ethnic groups. Finally, the Law Scholastic Assessment Test (LSAT) scores are equally predictive for minority and White matriculates, both in terms of grades and bar passage rates. This finding runs counter to the oft-expressed belief that test scores are biased against minorities; in fact, they are equally predictive for all groups.

In follow-up analyses in the next article, Detterman found that the difference in rate of passing the bar exam between those admitted by the regression model of LSAT and undergraduate grade point average (UGPA), and those not admitted by the model, showed a multiple *R* of .91, with the LSAT and UGPA for each minority group. This means that the poorer the minority student scores in the student pool, the greater the difference in law school passage rates. So scores do appear to influence minority students' pass rates.

Detterman's fundamental question concerns whether the United States wants a meritocracy that either does or does not limit access to some segment of the population. If we base a meritocracy on test scores, then we will certainly restrict racial diversity. Detterman claims that if we have open admissions, then we will still have a meritocracy but without initial restrictions on diversity. In this scenario, he argues that colleges would be forced to provide educational programs that actually educate people.

Wightman's 1997 article can be invoked to make a strong case for open admissions of the type that Detterman proposes. The degree to which this would work in practice would depend on the extent of our commitment to diversity, as well as what we as a society are willing to accept in order to achieve diversity. As Detterman reminds us, a superior approach would be to find out why there are racial differences and correct the problems that can be corrected. Regardless, if you agree with his analysis of the situation or with his many challenges of cherished assumptions (e.g., the value of being educated at elite institutions), then Detterman's reasoning is a breath of fresh air. His article deserves to be read by every policymaker and educator who is committed to AA. It will disturb some, but his arguments must be addressed.

In contrast to Detterman's position, Diane Halpern argues that if the goal of higher education is to select individuals most able to benefit from their offerings to become "successful citizens," then we run into the problem of defining what we mean by successful. Halpern asks, Is being a good parent as successful as being a good politician or earning a lot of money? Obviously, colleges could fall back on the claim that they use more immediate and modest criteria, such as the grade point average (GPA), that are objective and "fair." But even here, Halpern urges a deeper nonstatistical definition of fairness: Is it fair to give a low aptitude score to someone who has not had the opportunity to learn? What if this individual had the inherent talent to score higher but lacked the opportunity to do so? Unlike Detterman, Halpern challenges the basic assumption that aptitude scores reflect aptitude, at least in any straightforward manner. Despite this, Halpern would not argue that minority applicants who had less opportunity to learn the content of the LSATs should nevertheless be admitted to the University of Texas Law School with index scores (on the basis of a linear combination of the LSAT and GPA) that would have precluded Whites from being admitted (see *Hopwood v. State of Texas,* 1996). No one wants to sacrifice educational standards even if he or she believes that aptitude scores are not good measures of native aptitude. Clearly some problems exist conceptually with standardized scores, but alternatives are not obvious.

Halpern offers a number of cogent cautions for readers unfamiliar with psychometrics, such as the reminder that validity depends as much on test content which can be arbitrary (e.g., the proportion of a mathematics aptitude test that comprises calculus problems v. geometry problems), as on ability. Test predictiveness can be altered in many ways, and validity may be improvable, but the only way to know for certain is to experiment with testing formats and contents.

Wightman comes to this special issue having published what is still the best quantitative study of the predictive validity of various models including the LSATs and UGPA. Although LSATs and UGPA account for three-quarters of the variance in law school admissions decisions of Whites, adhering to the same statistical models would devastate minority admissions to levels not seen in nearly 40 years. The problem, as Wightman shows, is not alleviated much by totally abandoning the LSAT; using only UGPA without the LSAT to make admission decisions would still reduce Black admissions by nearly two thirds. A stark contrast exists in her data: The mean LSAT scores of Black and Hispanic applicants from the highest socioeconomic group are lower than the mean LSATs of Whites from the lowest SES group. Thus, using SES instead of race would not solve the diversity problem in law schools either, because racial differences persist within the same socioeconomic group.

Yet, Wightman shows that factors other than LSATs and GPA must have been in use for the final quarter of the White matriculants in law school because they would not have been admitted on the basis of a straight application of the linear model of LSAT and GPA (conversely, a much smaller number of high-scoring Whites were denied admission).

Along with the article by Julian Stanley (2000), which I comment on later, Perloff and Bryant's article is the most fun to read, not only because these scholars are irreverently straightforward with their language and willingness to confront sacred cows, but also because they go beyond the extant data to advocate a plan that may or may not achieve their objective. They see the possible pitfalls but are willing to accept that price. Before I comment on their proposal to increase diversity in schools and workplaces, I want to comment on the premise of their argument; namely, that diversity, not AA per se, is the source of payoffs for industry and institutions of higher education.

Why is diversity valued as an end in itself? Perloff and Bryant rely on personal testimonials of business and education leaders for their support. For example, Albert Carnesale (quoted in Lewis, 1997), the president of the University of California, Los Angeles, asserts, "education is markedly advanced by diversity—because students learn from each other," and Nannerl Keohane, the president of Duke University (quoted in Higginbotham, 1998), writes that her "experience as a teacher at three institutions of higher learning and as the president of two others is that diversity benefits students, faculty, institutions, and the world of knowledge."

But personal testimonials are just that, and it is fair to want more hard empirical data showing that diversity in the classroom really adds to the educational experience of students of all groups. Are there such data? Perloff and Bryant inform us there are not and there are not likely to be any forthcoming. I think they are correct. The smuggled assumption in the testimonials seems to be that diversity is good because students benefit from being exposed to contrary points of view. Yet universities are not notably open to contrary points of view if these alternative views go against popular values and beliefs. Consider how many women's studies programs strive to achieve diversity by recruiting scholars who are antiabortion or how many Middle Eastern studies programs try to diversify their faculty by hiring those with pro-Palestinian views. Admissions officers are not known for their zeal in recruiting Appalachian Whites with histories of deprivation, even though their presence in classrooms and dormitories could be informative. In short, diversity is endorsed as a goal only when it comports with or fosters current values and beliefs—the very antithesis of what it is supposed to be.

A related assumption in the prodiversity argument seems (to me, at least) to be that students will be more likely to envision different points of view and test alternative hypotheses to their pet theses if they are exposed to others who harbor different views; but is this really true? Aren't we in the business of challenging students to think critically? To pose plausible alternative hypotheses? Do we expect students, once they leave the classroom to require the presence of different views in order to think about them? I hope not.

My questioning of the premise that diversity is good is not meant to suggest that I find fault with Perloff and Bryant's reasoning about race-conscious selection procedures or about the payoffs for racially diverse workplaces and classrooms. Indeed, it is precisely because these authors are so lucid in their writing and thinking that I find myself able to raise these issues at all. Perloff and Bryant are correct: Societies that are as diverse as ours can expect significant pay-offs if the ethnicity of police reflects that of the communities they serve, or by having marketing analysts who are aware of, and identify with, the targeted market. Perloff and Bryant deserve credit for making such pay-offs explicit. Where I quarrel with them is with the assumption that diversity ineluctably leads to pay-offs. I can think of many diverse views (and the individuals who espouse them) that I would not want my students or child exposed to. On the other hand, I lament the real lack of earnestness on the part of educators when it comes to recruiting the kinds of diversity that could expand our students' cognitive and emotional horizons.

Finally, let me comment on Perloff and Bryant's proposal to create separate ethnic databases of applicants and normatively standardize raw scores on a variety of variables deemed to be relevant for a particular school or job (e.g., test scores, work samples, GPA, extracurricular activities, and letters of reference). Like Wightman, they bemoan the need for better and more meaningful measures for admissions and hiring decisions. Armed with such measures, Perloff and Bryant could then, for example, employ regression analyses to establish weights to optimize prediction of grades or work efficiency within each ethnic group.

It seems to me that this scheme entails some brand new positive aspects (e.g., race would be linked explicitly to job needs or educational needs and selection would preserve merit-based rankings within racial group) but the scheme also includes some of the politically unpopular aspects that race-norming and quotas entail. I think Perloff and Bryant's plan would work only if it could be shown that it possessed more validity for school or job performance than the current procedures (mainly the use of standardized test scores, GPA, interviews, and letters of reference). This may not be hard to do given the limitations of such procedures. The onus, however, would fall on Perloff and Bryant to demonstrate that their system, which is really designed to preordain a specific number of minorities in any given environment, works as they anticipate. Putting aside my numerous questions about their proposal, Perloff and Bryant are to be applauded for coming up with an original and forthright plan that could help solve one of the thorniest problems facing us.

In the study by Matthew Scullin and his associates (Scullin et al., 2000), there is evidence that Black students' cognitive test scores measured in high school are underestimates of their ultimate cognitive attainments at the end of college. Other researchers have shown a similar effect, suggesting that the Black high-school experience is not as academically challenging as the high-school experience of many Whites. The result is that when Blacks are put in a challenging setting (college), their test scores rise. This is good news because it demonstrates the need to take high school scores with a grain of salt if we are interested in longer-term outcomes. They may not be predictive. This is exacerbated by Scullin et al.'s finding that educational attainment is a better predictor of Blacks' progress than IQ. Although Scullin et al. do not delve into the reasons for this finding opposite arguments can be put forward including both the discriminatory role of tests for Blacks and the Gottfredson (2000) argument, namely, that employers use educational attainment as a screen for cognitive status. If the latter is correct, then hiring African Americans with lower test scores than Whites at comparable educational levels could end up signaling employers that the variable is not working as assumed. Scullin et al. also provide evidence that, on average, Blacks are remunerated in excess of their test scores and familial backgrounds. Controlling for cognitive scores shows that White women are the most undervalued group, earning significantly less than the other groups. It will be important to determine whether this is true at all ability levels or just at the mean.

In the final article in this section, Everson reviews recent advances in cognitive science. If future aptitude tests can be rooted in our emergent understanding of the way in which information is processed, then there is hope that new forms of assessment can be developed to identify types of talent not presently recognized. Everson holds out the hope that such a testing initiative may be useful for recruiting minority students and others who fare poorly on tests that are saturated with the general factor.

Perhaps the single most exciting initiative, if it can be made to work, is the integration of cognitive psychology's insights about how students process knowledge into model-based assessments. If demonstrated to be practical, then such model-based measurements could, as Everson himself points out, transcend the traditional ranking function and link assessment to actual instructional plans. This would be an enormous advance.

Four aspects of Everson's vision that are not spelled out, however, concern: (a) the commitment of the College Board to develop new measurement models that reflect a learner's actual knowledge representations, strategies for learning, and metaknowledge; (b) the willingness of elite universities and their faculty to change the way they admit and teach students in response to feedback from new, diagnostic testing; (c) the likelihood that model-based measurement will provide detailed pedagogical information that will enable instruction to be individualized; and (d) the willingness of elite universities to accept as part of their mission a different type of teaching program, one geared to each individual student's processing characteristics. If the responses to these issues are affirmative, then concerns about diversity and equity could become moot.

Part III: Intelligence as Process

In the first article in this section, Deary, Austin, and Caryl (2000) make the useful distinction between measurement–prediction, on the one hand, and explanation, on the other. Rather than proceeding from theory and explanation to measurement, intelligence testing proceeded as though the two processes were independent. The result is that we now have forms of measurement that boast high predictive validity, but we are not sure why. We do not know what, if any, process links the measured performance to relevant biological processes needed for intelligent behavior.

Deary, Austin, and Caryl hold out hope that psychologists and biologists will some day make progress on the explanatory level, though they are appropriately gloomy that this will occur in the near future. As one example of the problems such progress will entail, these authors cite Chorney et al.'s (1998) recent demonstration that an insulin-fixing gene on the long arm of Chromosome 6 is associated with intelligence differences in children. Although this could be seen as a first step at a biological explanation of mental test scores, it presents barriers that may characterize all such efforts. For starters, there is considerable overlap in the presence of this genetic marker in the blood of high- and average-intelligence children. About one third of average-IQ children actually have a greater association with the presence of this marker than do the high-IQ children. Most biological markers are probabilistic and likely to be relativistic; moreover, no compelling theory exists that ties this genetic location to cognitive differences. It was not a theory that motivated Chorney and colleagues to research this genetic location as much as it was the opportunistic advantage of the greater ease offered by investigating this location.

In the second article, Robert J. Sternberg demonstrates once again why he is regarded as the premier researcher in the field of intelligence. His vision is a constantly questioning one that spans psychometric and contextual approaches to intelligence and moves with ease across educational, social, and cognitive psychology.

Sternberg (2000) raises the question of circularity in his article. No matter how we attempt to validate the construct of intelligence, warns Sternberg, there are conceptual snafus. Parts of this argument are not new, such as his point that all predictive systems are social inventions rather than biological dictates of nature (see also Halpern's article in this issue for a related argument). However, Sternberg weaves the parts of his argument into a novel framework that prods the reader to see instantly the pitfalls of all current IQ validation schemes. His demonstration that scores on both crystallized (e.g., vocabulary) and fluid (e.g., mazes and matrices) tests of intelligence are at times uncorrelated, or even negatively correlated (as shown by his work with Grigorenko in Kenya), with measures of social and physical adaptation, is shocking. We have all been inundated with data showing moderate correlations between intelligence and job and school success, and yet Sternberg reminds us that this is a culture-specific phenomenon; change the cultural setting and the correlations disappear or even reverse. As interesting as this "now-you-see-it-now-you-don't" argument is, it is Sternberg's larger point that will probably lead to this article being frequently cited; namely, that the stories of successful role models that surround us will animate very different types of behaviors. At times those behaviors will be consistent with doing well in school and on IQ tests (and in such cases the validity coefficients will be respectable), and at other times the behavior will militate against doing well on academic and psychometric tasks (and consequently will lead to low validity coefficients).

One issue that was not clear upon reading Sternberg's provocative and thoughtful argument concerns the construct validation enterprise that has convinced many in the psychometric camp of the reality of "intelligence." Sternberg does not touch on this enterprise other than to critique the way school grades and job success have been used to calculate validity coefficients. But what about the interrelationships between IQ scores and a host of biological measures (head circumference, cranial blood flow, heritability, nerve conductance velocity and oscillation)? The fact that IQ seems to be tapping into some biological substrate seems, to some, to give its correlation with school grades and job success a veneer of validity: Good nervous system efficiency is important for both school and work success as well as for scoring well on an IQ test.

None of this is news to Sternberg; he knows all about these interrelationships. In the future it will be interesting if he tackles this issue head on, and explains in quantitative terms the extent to which his contextual subtheory can interact with the other components of his triarchic theory to produce mismatches between test performance and school performance in some situations and, yet, be consistent with the correlations with biological measures. Mathematically, there is room for this to occur but it will be interesting to see a worked example.

The third article in this section, by Joseph Fagan (2000), argues that intelligence should be construed as processing efficiency rather than as processing knowledge. Fagan's own Infant Intelligence Test is an example of what is meant by processing, as it depends on the infant's differential response to novel stimuli, hence, reflecting short-term memory ability. Processing, as Fagan shows, is related to IQ scores but not to racial differences (because there are none when it comes to processing on his task).

Construing intelligence as processing is a provocative proposal; it allows for early detection of disabilities that predate decrements on later IQ; it allows for the principled distinction between children whose processing is normal but who have other problems, and it permits a range of disability assessments to be made that are currently hopelessly comingled with IQ scores.

If intelligence is reducible to processing, then what about other measures of processing such as choice reaction time? (See Jensen, this issue.) Fagan is aware of this work but argues that Jensen's choice data reveal no racial differences on the least complex tasks (i.e., those that are presumably more basic indexes of underlying processing and are uncontaminated by cultural influences), but do reveal differences on the most complex tasks, presumably because of the influence of cultural factors on the latter. However, it is not obvious why or how culture could exert influence on such tasks, given their simplicity and the fact that not all studies of schooling have reported an influence on short-term memory performance. Along the same lines, it would be interesting to test Fagan's argument about the primary role of processing by examining other measures of processing such as visual inspection (see Deary, Austin, & Caryl, 2000). What if some measures of processing reveal different patterns of results than are revealed by tasks involving novelty preference?

Fagan's argument is, at present, a promissory note but its promise is so important for society that future research could be very fruitful. By shifting the discourse from knowledge-based measures of intelligence to processing measures, Fagan has done much more than merely switch the rhetoric. He has offered a fundamental change in the way we conceive of and assess ability.

Part IV: The Dilemma of Group Differences

Articles by Arthur R. Jensen (2000), Linda S. Gottfredson (2000), Cecil Reynolds (2000), and John Hunter and Frank Schmidt (2000) are all firmly rooted in the psychometric tradition, and these authors are among the most respected in this tradition.

Jensen outlines the basic psychometric validation enterprise quite well, drawing on the basic research described in his impressive recent book, *The g Factor* (Jensen, 1998). He makes a very important point that is often lost on policymakers, so let me make it explicit here: Jensen argues that the use of any test that has validity in predicting success in schools and universities will result in racial asymmetries that are similar (though not exact) to those that result from the use of IQ tests and their surrogates (SAT, GRE, and ACT). So, if we abandoned the use of the SAT for college admissions, and if a college was highly selective and needed some principled basis for deciding which of their many applicants they wished to offer places to, then the use of high-school grades or tests that have high face validity (e.g., tests of math and verbal comprehension that are chosen after consultation with college authorities to reflect the skills needed to succeed in college), would also yield racial imbalances. This is because any such tests are heavily *g* loaded—the same factor that is responsible for the predictiveness of the IQ and SAT. *G* can be derived from any correlation matrix that is composed of diverse tests (math, science, reading, and spatial reasoning, etc.), and it is the single most predictive ingredient, more so than specific factors, for most jobs and colleges.

Why do I single out this one argument? Because in putting it forward, Jensen rightly challenges critics of tests to come up with an alternative to SATs and IQ tests that will not result in adverse impact on racial minorities. Even tests that are not called intelligence tests and are more closely tied to what colleges expect students to master, will, according to Jensen, result in disproportionately fewer minority admissions.

My sole quarrel with Jensen's argument has to do with his claim that early intervention is ineffective as far as raising IQ goes. In his words:

> No method of psychological or educational intervention has yet demonstrated reliably the power to make sizeable or enduring upward changes in children's IQ, or particularly their level of *g* . . . . [P]sychologists and educators do not know of any means for raising the *g* level of children who are at risk . . . . The largest authentic gains in *g* that have been induced experimentally in children by the most intensive and extensive means ever tried amount to, at most, about one third of a standard deviation (equivalent to 5 IQ points), and it is not yet known if this amount of gain will last to maturity. (p. 125)

Although it may be true that the verdict is still out insofar as the durability of IQ gains is concerned, there is solid evidence that early intervention can lead to large IQ gains. Craig Ramey and Sharon Landesman Ramey (1998) review this evidence for their own multisite clinical trials and report IQ gains on the order of 13 points for 3-year-olds. That is, children exposed to intensive early enrichment programs outperform their unexposed peers by around 13 IQ points at age 3 years. Will such increments last? Jensen may be right in thinking they will not. Only time will tell, of course.

The second article in this section is Gottfredson's (2000), who begins where Jensen leaves off. I confine my comments to Gottfredson's most striking assertions although there are others that deserve to be commented on when greater space permits.

At the outset I want to make clear that my disagreement with several of her claims is a purely scientific one; Gottfredson's article is far more data-based and thoughtful than the vast majority of her critics, and I wish the latter would attempt to justify their positions with the same emphasis on systematic data that Gottfredson uses to justify her own position. She is correct in condemning many of us who cannot refute the three-pronged empirical phalanx of data she reviews—that the racial skills gap is (a) real, (b) stubborn, and (c) important—but who instead seize on "wisps of evidence that can be construed to contradict one or more of the three conclusions" (p. 139). Pecking at the fringes of one or more pieces of evidence is the stuff that wins debates but loses the argument in the estimation of fair-minded folks who are familiar with the entire corpus of data, as is Gottfredson.

I am in agreement with Gottfredson's claims about the magnitude of racial differences in intelligence scores. They are quite large. Blacks score around 1 *SD* (15 points) lower than Whites on IQ tests. However, I am less convinced that this 1 *SD* gap in IQ translates, at least in any linear manner, into a 1 *SD* gap in job-related skills. The predictive validity of IQ is only modest for most jobs, perhaps between .3–.4. Thus, a skills gap of this magnitude seems more likely. This is not in itself a solution but it does reduce the racial disproportionality of the relevant job applicant pool. Also, Granted, that although Gottfredson is right about the racial gap in mental test scores being quite stubborn (it has existed ever since the first mass administrations of such tests in the early 1920s), there is some evidence that the mental-test score gap is far from immutable. Elsewhere, I and others have shown that the racial gap among some cohorts narrowed by between one-third and one-half in the years between 1971 and 1988 (Ceci, Rosenblum, & Kumpf, 1997; Grissmer, Williamson, Kirby, & Berends, 1997; Williams & Ceci, 1997). Thus, although I am not disputing any of Gottfredson's three prongs, I am suggesting that the interpretation of two of them is open for debate.

Putting aside the validity of these points entirely, however, I find it hard to proceed from them to Gottfredson's position that schools and employers will be faced with a no-win situation because the reality of insisting on racial proportionality will result in disillusionment and resentment. Gottfredson argues that, if universities and employers admit or hire a proportionate number of Blacks, then these individuals' low-skill levels will necessitate huge dyseconomies in training because Blacks will be, on average, an entire skill (1 *SD*) below Whites. It would thus take longer to train or educate Blacks and they would require different types of training than Whites. However, it seems to me that the alternative—a lack of proportionality in admissions and hiring—also breeds resentment. Departures from parity in the workplace (e.g., racial asymmetries in hiring and promotion) are taken as prima facie evidence of discrimination regardless of whether they are justifiable on the basis of evaluation. Politically, both of these situations are untenable. In the near term, this means that we should expect a continuation of fury over AA because "we can't live with it or without it." Perhaps if we heed Gottfredson's advice, then programs to ameliorate the skill gap will work their magic in a generation or two. Even a closing of the gap by one third to one half, as was witnessed in the 1970s and 1980s, would go a long way toward reducing the dilemma.

The next article in this section is by Cecil Reynolds (2000). He quite understandably laments the tendency not only of the media but even of psychologists (who should know better) to make accusations about racial bias in mental testing. Reynolds points out that bias is a complicated concept. He notes that it is a very different matter to ask if a professional can identify test questions that are insulting to a given racial group than it is to ask if a professional can identify questions that are psychometrically unfair. The latter has little to do with the existence of racial disparity in test scores; that is, evidence of bias cannot be demonstrated because a test is associated with mean differences in racial groups any more than the owners of National Basketball Association (NBA) teams can demonstrate evidence of bias because of racial disparities in team makeup. Reynolds argues that professional psychologists are notably inaccurate when they are asked to identify test questions that are racially unfair (i.e., likely to be more difficult for a given racial group because of unfamiliarity). Similarly, it is lamentable that professionals assert that one test is less biased than another because it happens to be nonverbal.

One possibility not explored by Reynolds is that a testing enterprise can be biased not because the predictor (test) underpredicts the capability of a given group but because the criteria (e.g., school or job success) require assumptions about the familiarity of the content of both the test (predictor) and criteria. One could argue that even if the test is psychometrically fair in the sense that it is equally predictive of school or work success for Blacks and Whites, it could nevertheless be an unfair enterprise. No one would ever agree that a test of current facility in speaking Russian is a fair predictor of future aptitude for Russian for someone who never attended a school where Russian was taught. Similarly, one could claim that tests might be developed that are better gauges of Blacks' aptitude than are the current SATs and IQ tests. Perhaps if such tests were developed, then they would be shown to be better predictors of Blacks' school and work performance. It is an argument that has nothing to do with statistical fairness, of course, but it may be what fans the flames of antitesting proponents.

In the final article of this section, Hunter and Schmidt extend Reynolds' argument to the level of specific test items. In their usual careful manner, these authors point out that claims of racial or gender bias, in particular, test questions are unsupported by the data when care is taken to avoid artifactual problems and the use of unwarranted assumptions. On the one hand, I always considered it highly improbable that although no bias exists at the level of the total test it nevertheless does exist at the level of individual questions. Although such a state of affairs is logically possible, it never seemed likely. Now Hunter and Schmidt explain why the handful of findings in the literature that have supported the bias argument are themselves flawed. Along with Reynolds' article, this paper ought to give pause to those who continue to make claims about test bias.

Part V: Individual Differences: Implications for Educational Policy

The lead article in this section is written by none other than James Flynn (2000), the man whose name is attached to the worldwide rise in IQ in this century, the so-called *Flynn effect.* Flynn has had an uninterrupted series of very important insights, and he has continued this series in the present article. His point is that because of the upward creep in IQ scores that takes place each time an IQ test is renormed, one can expect to see a gradual, steady inflation of IQ scores until the next renorming occurs, 15–20 years later. To put this concretely, Flynn argues that there may be as much as a 7-IQ-point increase over a single renorming cycle, say between 1972 and 1989. So, a child whose performance earned her an IQ of 65 in 1974 would be expected to earn an IQ of 72 in 1989 as a result of the upward creep known as the Flynn effect. This represents an enormous increase because it could change the child's status from mental retardation (MR) to regular education (RE), as some local educational authorities used a cutoff score of 70 to determine eligibility for MR classes. [A score of 70 would be the cutoff if the Wechsler Intelligence Scale for Children—Revised (WISC–R) was used, as it has an *SD* of 15, thus allowing 70 to mark an *SD* cutoff.]

On reading Flynn's article, I immediately saw the implications of what he wrote for another special population, the gifted or talented. In this case, the argument works as follows: If a child gets an IQ of 128 in 1974, then she would be expected to score higher as that set of IQ norms gets older, so that by 1989 she might score around 135—high enough to qualify her for classes for gifted and talented. The Flynn effect works in reverse, too. Imagine how distraught parents would be to learn that their child, who scored 135—and was deemed eligible for gifted classes in 1973, was tested a year later when the new norms were in force, and now scored 7 points lower—thus, no longer qualifying for gifted or talented programs. Heinous!

I decided to test one of Flynn's predictions by examining data posted by the Department of Education on the numbers of children classified as receiving services for mental retardation between the onset of the new norms for the WISC–R in 1972 (but not made widely available until 1974) and its successor norm, the Wechsler Intelligence Scale for Children-III (WISC–III), published in 1989. It turns out that the data prior to 1977 were not available, so I constrained my analysis to the data covering the period 1977–1995.

Flynn's prediction would lead us to expect a dramatic decline in the number of children classified as MR between 1977 and 1989, then, an upward surge after 1989 when the new norms came into play. To some extent, this pattern is what happened. The number of MR classifications in 1977 was 960,000, and this number plummeted to 570,000 in 1989, the final year before the new norms were widely available. However, it was not the case that the number of MR classifications immediately rebounded to 960,000 in 1990—the first full year of the new norms' availability. Instead, there was a continuation in the decline in numbers for several more years, followed by a slight rise in 1993. Since that time, the rise has continued, though not dramatically. So, why wasn't the full impact of the Flynn Effect observed, if the upward creep in IQs is as steady as mentioned? There are several possible reasons.

Perhaps countervailing forces were at work to dampen the full force of the Flynn effect, such as changing policies regarding classification of MR (e.g., changes in the cutoff IQ score used, or the inclusion of new criteria), or perhaps there was a sea change in parental attitudes about the desirability of the MR classification (and the services that go with it), or maybe there was a reluctance on the part of local educational authorities to adopt the new IQ norms as soon as they became available. This last possibility seems likely because test forms are expensive and many school psychologists do not immediately throw away old forms just because a "new and improved" test norm is available, especially when they are similar to the old forms.

One other reason that the number of MR classifications did not rebound to the 960,000 number may be a tendency on the part of school psychologists to adjust the new norms upward. That is, once they recognized that the new norms were harder than the ones they had been using for the past decade or more, leading to a sudden drop in IQs, they added IQ points to borderline children's scores by either scoring their tests more leniently, or giving them extra time to compensate for the harder norms. This is surmise on my part but it does accord with my observations and discussions with school psychologists.

Regardless of the explanation for the above-mentioned data, Flynn has demonstrated once again his capacity for making creative and important insights. I had not considered these possibilities until I read his article. Now, I am eager to learn more about the impact of the Flynn Effect on the lives of individual children, including those whose classifications could have been changed had they been tested with different norms, and, in fact, I am pursuing research on these issues. For this, I have Flynn to thank.

In the second article in this section, John Bishop (2000), a colleague at Cornell, reports on his provocative findings on the use of curriculum-based exit exams. When I first heard him report these findings, I urged the editor of this special issue to invite him to write about them. I am glad she did. These are the kind of data that policymakers should use to forge policy. Bishop shows in his careful analyses that the use of exit exams can be expected to raise performance, irrespective of social class, school spending, and other demographics. Few pieces of social science research have the import that Bishop's has, and I would hope that Bishop's findings will serve to spur a national debate. He argues persuasively that none of the old excuses (e.g., large numbers of students speaking a foreign language) can be used to justify low performance.

The second article in this section is Julian Stanley's (2000). Previously, I noted that this article is a lot of fun to read, for Stanley regales the reader with example after example of the logic and common sense that has been associated with his name for the past half century. Who, after reading his stories about high-achieving students let down by the school system, can argue that the present educational practice of treating all students in the same age-graded, lock-step manner is in their best interests? Not I. Stanley's is a voice of reason in the often acrimonious debate over ability tracking. He reminds us that the problems we are witnessing in this debate are self-inflicted: Why don't we shift from A–F grading to statements about each student's achievement level?

I think that even the most extreme critics of tracking would accept Stanley's proposal for differential teaching of students on the basis of diagnostic testing if it were stated as follows: Don't allow students who already know most of what the teacher plans to teach for the entire school year languish in an environment where they will become increasingly bored, and actually end up at the conclusion of the year scoring worse on a standardized achievement test than they could have scored at the beginning of the year. Note that I have not used the words precocious or tracking. Neither have I done something else: I have not cast Stanley's rationale in terms of allowing students at a given ability level to be taught at their own accelerated pace so they can end up achieving far more of their potential than they would if they were forced to take classes with everyone else. This latter framing of the rationale arouses the ire of parents, politicians, and school board members because it shifts the argument away from cruelly dampening a student's knowledge and motivation by forcing him or her to spend an entire school year reviewing what is already known at the outset, to fostering the growth of individual talent so that some students can flower in an accelerated environment while others languish in RE. Words are weapons in the debate over tracking and Stanley can convince us if he avoids certain words. Of course, this is deceptive; hence, the reason he does not use certain words. He is aware that the problem is not only one of preventing children from repeating a year of instruction that they already know, but it is also one of allowing students who are able to learn more to do so even if they do not already know all that they will be taught in a regular classroom. Thus, some degree of acceleration is entailed.

I wish that school board members would read Stanley's article and defend their resistance to allowing children to take classes with older students who are at their level in a given topic area. In the media coverage of this debate, I have never heard responses to the kind of examples that Stanley gives, yet we know that such children exist, and in nontrivial numbers, too. I hope Stanley lives to see the revolution in diagnostic teaching that he urges us to adopt.

The final article in this section is David Grissmer's (2000). Unlike the articles by Stanley and others who focus on the use of tests like the SAT to make diagnostic and admission decisions for individual students, Grissmer is concerned about the misuse of SAT data when it is aggregated and reported in the media. As he points out, aggregated SAT data is used by local educational authorities to argue for more funding, by politicians to blast the job the schools are doing, and even by realtors to sell properties in high-scoring school districts to families with school-aged children. If the aggregated data were valid, then there would be no problem, but as Grissmer shows, there are significant problems with reporting scores in this manner. Changes in SATs may have nothing at all to do with changes in educational quality. Demographic changes in the makeup of students taking the SAT can cause large fluctuations in mean scores (e.g., increasing numbers of minority students are taking the SAT and their scores tend to be lower than White and Asian-American students' scores), as can changes in trends to apply to an in-state college that does not require the SAT versus an out-of-state college that does require it. For example, there have been dramatic shifts in the past 20 years in the proportion of colleges that require the SAT; this has resulted in some of the most select students taking the SATs because they wish to attend elite out-of-state colleges. At the same time, their in-state college requires the ACT test. To whatever degree the distribution of colleges requiring the SATs versus ACTs has changed over time, this, too, will confound the interpretation of aggregated SAT data.

Using the SAT to diagnose the job school admissions personnel are doing is a little like using survivorship–mortality data to determine which hospital department personnel are doing the best job to promote health. Imagine the reaction of those in the oncology department upon learning they are doing the worst job in the hospital because more of their patients are dying than in, say, the patients in the maternity ward. The oncology department personnel could rightly claim that they are dealing with far more grievous problems. Grissmer is making a related point, I think: For example, imagine that high-school teachers are doing an increasingly better job teaching students who have various kinds of disadvantages (learning disabilities, poverty backgrounds, and emotional problems), and as a result of this better job more of these disadvantaged students graduate and wish to apply to college. If this were to happen, then the mean SAT score for this high school would decline because these disadvantaged students would, as a group, score lower on the SAT, thus dragging down the school's mean.

Grissmer argues that none of the interpretative snarls that inhere with using mean SAT data are inevitable. Schools could use the National Assessment of Educational Progress (NAEP) data instead. The NAEP is a far superior tool for making inferences about the job that school personnel are doing because (a) it is based on a nationally representative sample of fourth, seventh, and twelfth graders, (b) it is rooted in curriculum, and, most important, (c) the test items have remained constant since 1971. So it is possible to actually use the NAEP to determine whether students today are getting as many answers correct as did their predecessors. (The SAT content is constantly changing, making such comparisons difficult, at best.) This is why Grissmer makes the point that between 1970 and 1990 the SATs were declining at the same time the NAEP scores were rising! The public emerged from this period with the impression that school personnel were doing a bad job. However, better data—the NAEP—lead to the opposite conclusion.

Assessing educational progress is a very difficult enterprise in the best of conditions, as measurement experts realize. It does not help matters when data as unsuitable to making comparisons, as are aggregated SATs, are used to make judgments about the effectiveness of the job being done by individual high schools, school districts, states, or nations. If I could wish for one message in this entire special issue to get the attention of the public, then it is this point of Grissmer's.

Conclusion

As should by now be obvious, there is considerable agreement about the usefulness of measures of general intelligence whenever admissions officers or personnel officers are forced to choose among a group of applicants that exceed the number of slots. Such decisions are informed by the inclusion of a measure of general intelligence. Even stubborn critics of testing seem to grant this.

However, it is an entirely different matter when it comes to the meaning of so-called general intelligence or why it improves prediction, or even if its predictiveness is grounds for its use. After all, mere prediction can be achieved in numerous ways that society would not find acceptable. Imagine being told that your household honesty, assessed by your answer to questions such as "Have you ever lied to your parent?" is predictive of workplace honesty. Even though the predictiveness of such a question could be established, we could argue that it ought not to be used for a variety of reasons. Similarly, both race and SES are predictive but society does not tolerate their explicit consideration, save when they are positively weighted as in Hopwood or Bakke. Even height is slightly correlated with IQ but if it could be demonstrated that it adds to the predictive equation for admission or hiring, then no one would advocate its use. My point is simply that a disjunction exists between prediction and explanation. When it comes to prediction, there is broad agreement that measures of general intelligence (the most popular of which is the IQ test itself and its surrogates such as the SAT and GRE), do improve predictions of grades and supervisor ratings. However, the explanation given to the predictions is still open for debate.

Many authors of articles in this special issue opine that general intelligence achieves its predictiveness because it is an index of abstract reasoning. For example, Gottfredson, Jensen, and Hunter and Schmidt have proffered this view. Yet, no one really knows what it means to engage in abstract reasoning. What one individual construes as abstract, is seen by another as concrete. Let me give a personal, unpublished, example to illustrate this.

In 1989, Narina Nightingale and I asked a group of eminent scholars to tell us how they defined abstract reasoning. These individuals, all eminent scholars who can claim to be working on deeply abstract levels in their own work, included a group of 20 Nobel Prize winners (most from physics, literature, and chemistry), as well as a group of individuals who were acknowledged to be truly outstanding by those working in their disciplines. Mathematicians in the survey were recipients of the highly coveted Sloan Award, and linguists and philosophers–logicians were members of the prestigious American Academy of Arts and Sciences; their peers had repeatedly nominated them in our survey as among the most brilliant scholars in their disciplines. We also queried past and present Poet Laureates, winners of Pulitzer Prizes for fiction, and a number of eminent psychologists, including two who had won Nobel Prizes themselves for contributions to decision theory and medicine (Herb Simon and Roger Sperry). As already mentioned, these nominees were a highly decorated lot, having won prizes of one type or another (Guggenheim, Nobel, Sloan, MacArthur, and Myamota).

We sent these eminent scholars a list of 32 questions and problems and asked them to rate these items on the basis of the level of abstraction required for their solution processes. We selected particular questions and problems for specific reasons. The questions and problems included those that, at the time, were associated with the largest Black–White differences, as well as those that were associated with the smallest differences, on the most commonly used IQ test (Wechsler–Revised). Altogether, these questions and problems were chosen on the basis of their psychometric properties including overall pass rates and canonical variates with Full Scale IQ (Sandoval, 1979, 1982; Jensen, 1977).

The survey results showed that not only did scholars from one field disagree with scholars from another field about the abstractness of the 32 items, but also within a given field of inquiry there was usually little consensus. For example, some respondents rated the question "What does brave mean?" as more abstract than the mathematical problem "A jacket that usually sells for $32 was on sale for 1/4 less. When no one bought it, the store owner reduced the sale price by 1/2. How much did the jacket sell for after the second price reduction?" Their judgment was that the latter problem was algorithmic and did not, therefore, require deep abstraction, whereas the meaning of "brave" did. Others, however, assigned an exact opposite rating to the two questions. Such lack of consensus was found for the entire set of items in the survey.

Interestingly, when Nightingale and I divided the items into those that were associated with the largest racial differences and those that were not, and then computed the mean ratings for each group, we found that there was no difference in abstractness for the two groups.

I think the moral of this abstractness-rating exercise is that, if there is no agreement on whether the level of abstractness involved in items associated with the largest racial differences is greater than the level of abstractness associated with the smallest differences, then how can we assert that abstractness is the basis for the reported racial differences in general intelligence? This result should serve as a caution to those who would dismiss the complex performance of low-IQ persons (if and when it occurs) as somehow being less abstract than it seems. Simultaneously, we ought not elevate the complex performance of high-IQ persons (again, if and when it occurs) on the grounds that high-IQ persons are more abstract in general.

Such experiments demonstrate that it is possible to evaluate empirically more of the claims in the testing–AA debate than are currently explorable by concentrating solely on existing archival data. The debate can be advanced only so far if it is rooted solely in test score trends, and some experimental manipulation will be necessary to test competing claims.

References

Bishop, J. H. (2000). Curriculum-based external exit exam systems: Do students learn more? How? *Psychology, Public Policy, and Law, 6,* 199–215.

Calvin, A. (2000). Use of standardized tests in admissions in postsecondary institutions of higher education. *Psychology, Public Policy, and Law, 6,* 20–32.

Ceci, S. J., Rosenblum, T. B., & Kumpf, M. (1997). The shrinking gap between high- and low-scoring groups. In U. Neisser (Ed.), *Intelligence on the rise* (pp. 287–302). Washington, DC: American Psychological Association.

Chorney, M. J., Chorney, K., Seese, N., Owen, M., Daniels, J., McGuffin, P., Thompson, L., Detterman, D. K., Benbow, C., Lubinski, D., Eley, T., & Plomin, R. (1998). A quantitative trait locus associated with cognitive ability in children. *Psychological Science, 9,* 1–8.

Deary, I. J., Austin, E. J., & Caryl, P. G. (2000). Testing versus understanding human intelligence. *Psychology, Public Policy, and Law, 6,* 180–190.

Detterman, D. K. (2000). Tests, affirmative action in university admissions, and the American way. *Psychology, Public Policy, and Law, 6,* 44–55.

Everson, H. T. (2000). A principled design framework for college admissions tests: An affirming research agenda. *Psychology, Public Policy, and Law, 6,* 112–120.

Fagan, J. F. III. (2000). A theory of intelligence as processing: Implications for society. *Psychology, Public Policy, and Law, 6,* 168–179.

Flynn, J. R. (2000). The hidden history of IQ and special education: Can the problems be solved? *Psychology, Public Policy, and Law, 6,* 191–198.

Gottfredson, L. S. (2000). Skills gaps, not tests, make racial proportionality impossible. *Psychology, Public Policy, and Law, 6,* 129–143.

Grissmer, D. W. (2000). The continuing use and misuse of SAT scores. *Psychology, Public Policy, and Law, 6,* 223–232.

Grissmer, D. W., Williamson, S., Kirby, S. N., & Berends, M. (1998). Explaining trends in NAEP achievement test scores. In U. Neisser (Ed.), *Intelligence on the rise.* Washington, DC: American Psychological Association.

Halpern, D. F. (2000). Validity, fairness, and group differences: Tough questions for selection testing. *Psychology, Public Policy, and Law, 6,* 56–62.

Herrnstein, R. J., & Murray, C. (1996). *The bell curve.* New York: Free Press.

Higginbotham, A. L. (1998, January 18). Breaking Thurgood Marshall's promise. *New York Times,* p. 28.

Hopwood v. State of Texas, 361 F. Supp. 551 (5th Cir., 1994), *cert. denied,* 518 U.S. 1033, 116 S. Ct. 2581 (1996).

Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law, 6,* 151–158.

Jensen, A. R. (1977). An examination of culture bias in the Wonderlic Personnel Test. *Intelligence, 1,* 51–64.

Jensen, A. R. (1998). *The g factor.* Westport, CT: Praeger.

Jensen, A. R. (2000). Testing: The dilemma of group differences. *Psychology, Public Policy, and Law, 6,* 121–127.

Lewis, A. (1997, December 2). Developing real diversity: Even in post–Affirmative Action America, there are ways to encourage minorities, especially in higher education. *Pittsburgh Post-Gazette,* p. A27.

Perloff, R., & Bryant, F. B. (2000). Identifying and measuring diversity's payoffs: Light at the end of the affirmative action tunnel. *Psychology, Public Policy, and Law, 6,* 101–111.

Ramey, C. R., & Ramey, S. L. (1998). Prevention of intellectual disabilities: Early interventions to improve cognitive development. *Preventive Medicine, 27,* 224–232.

Reynolds, C. R. (2000). Why is psychometric research on bias in mental testing so often ignored? *Psychology, Public Policy, and Law, 6,* 144–150.

Sandoval, J. (1979). The WISC-R and internal evidence of test bias with minority groups. *Journal of Consulting and Clinical Psychology, 47,* 919–927.

Sandoval, J. (1982). The WISC-R factoral validity for minority groups and Spearman's hypothesis. *Journal of School Psychology, 20,* 198–204.

Scullin, M. H., Peters, E., Williams, W. M., & Ceci, S. J. (2000). The role of IQ and education in predicting later labor market outcomes: Implications for affirmative action. *Psychology, Public Policy, and Law, 6,* 63–89.

Stanley, J. C. (2000). Helping students learn only what they don't already know. *Psychology, Public Policy, and Law, 6,* 216–222.

Sternberg, R. J. (2000). Implicit theories of intelligence as exemplar stories of success: Why intelligence test validity is in the eye of the beholder. *Psychology, Public Policy, and Law, 6,* 159–167.

White, S. H. (2000). Conceptual foundations of IQ testing. *Psychology, Public Policy, and Law, 6,* 33–43.

Wightman, L. F. (1997). The threat to diversity in legal education: An empirical analysis of the consequences of abandoning race as a factor in law school admission decisions. *New York Law Review, 72,* 1–53.

Wightman, L. F. (2000). The role of standardized admission tests in the debate about merit, academic standards, and affirmative action. *Psychology, Public Policy, and Law, 6,* 90–100.

Williams, W. M. (2000). Perspectives on intelligence testing, affirmative action, and educational policy. *Psychology, Public Policy, and Law, 6,* 5–19.

Williams, W. M., & Ceci, S. J. (1997). Are Americans becoming more or less alike? Trends in race, class, and ability differences in intelligence. *American Psychologist, 52,* 1226–1235.