



Research Memorandum

ETS RM-12-07

Testing: It's Not Just Psychometrics

Lawrence J. Stricker

April 2012

Testing: It's Not Just Psychometrics

Lawrence J. Stricker
ETS, Princeton, New Jersey

April 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

<http://www.ets.org/research/contact.html>

Technical Review Editor: Daniel Eignor

Technical Reviewers: Nathan Kogan and Donald Powers

Copyright © 2012 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, GRE, and LISTENING. LEARNING. LEADING., are registered trademarks of Educational Testing Service (ETS).

ADVANCED PLACEMENT and SAT are registered trademarks of the College Board.



Abstract

This paper discusses research on the social and intrapersonal influences impinging on test-takers. One aim was to describe investigations of attitudes about standardized tests, particularly surveys of test-takers, the general public, and admissions officers concerning their attitudes about the validity and unbiasedness of college admissions tests. The other aim was to delineate inquiries into the impact of stereotype threat on test performance, focusing on experiments that assess the effects of Obama's presidential campaign on Black test-takers.

Key words: attitudes, admissions tests, stereotype threat, Obama effect, *GRE*[®] General Test

Acknowledgments

This is an invited address to the Division of Evaluation, Measurement, and Statistics at the American Psychological Association meeting, San Diego, CA, August, 2010.

Thanks are due to Brent Bridgeman, Ida Lawrence, and Cathy Wendler for encouraging this research on the Obama effect; Jackie Briel and Aleta Voda for providing the GRE General Test data for this research; Jason Wagner for tracking down surveys in the ETS archives; Fred Cline for preparing the figures; and Yigal Attali and Nathan Kogan for reviewing a draft of this paper.

Without doubt, psychometrics and statistics are at the center of the testing enterprise. Advances in these methods have revolutionized testing in the last half century. Consider these developments: item response theory, structural equation modeling, computer administered and adaptive testing, differential item functioning, statistical models of test bias, computer scoring of essays and other free-response tests, generalizability theory, and reformulations of validity theory.

But there is growing, though still fledgling, interest in the social and intrapersonal influences that impinge on the person taking a test. Test anxiety has been extensively studied for a long time, spawning a cottage industry with its own journals, book series, and scientific organizations, but other influences are only now being studied with any seriousness.

My objective is to describe two loosely related research projects in this area. My primary goal is to increase the testing community's awareness of the importance of these phenomena. A secondary goal is to point out both the value and the feasibility of using real-life data drawn from archives to address scientific and socially important questions.

Attitudes About Standardized Tests

The first project concerns attitudes about standardized tests. Test-takers' acceptance of the tests may affect their test performance, sapping it if they believe the tests are irrelevant or unfair, and thereby introducing extraneous variance that attenuates the tests' validity. Just as important, this acceptance—or lack of it—may affect test-takers' views of the institutions that require the tests. Believing the tests are irrelevant or unfair can poison test-takers' perceptions of the schools or employers that imposed the tests on them. The importance of test-takers' attitudes has been emphasized in recent years by a number of industrial and organizational psychologists interested in selection testing, beginning with Herriott (1989) and Iles and Robertson (1989). However, this issue has largely been ignored by other test users, with the notable exception of Nevo (1993), an educational psychologist concerned with school and admissions tests. Acceptance by test-takers, test users, and the general public is critical to the continued use of testing for hiring, school admissions, and other real-life decision-making.

Several papers, dating from the 1980s, discussing unpublished findings from surveys of attitudes about standardized tests made me aware of the often unrecognized value of such data. The first of these publications, by Lerner (1980), summarized findings through the 1970s. A later article, by Phelps (1998), and his recent follow-up chapter (Phelps, 2005) dealt with subsequent surveys. Stimulated especially by Phelps' yeoman efforts, I have taken a systematic look at

nationwide surveys of attitudes about admissions tests. Attitudes about admissions tests are of special interest because of the tests' ubiquity and the heightened public debate about their merits, chronicled in recent books by Zwick (2002) and Phelps (2003).

In order to locate unpublished surveys, I searched ETS's archives and got in touch with other major test publishers: the College Board, ACT, Law School Admission Council, and American Association of Medical Colleges (publisher of the Medical College Admission Test). I checked three major archives for survey data: the Roper Center for Public Opinion Research at the University of Connecticut; the Inter-University Consortium for Political and Social Research at the University of Michigan; and Polling the Nation, a commercial organization. I also searched PsycINFO for published surveys.

By this process I located more than three dozen surveys about admissions tests for undergraduate, graduate, and professional schools. Many of the surveys were commissioned by ETS. The surveys covered three broad topics: whether the tests are valid, whether they are useful, and whether they are unbiased. The respondents were test-takers and other students, parents of test-takers and students, the general public, and admissions officers and other school administrators.

I will discuss only a subset of these studies: those concerning attitudes of test-takers, the public, and college admissions officers about the validity and unbiasedness of college admissions tests. Although some of the data sets are accessible for reanalysis, I will be using only the statistics that are already available. With few exceptions, the questions and their response options differ from survey to survey, sometimes substantially, precluding precise comparisons of the level of attitudes.

Keep in mind also that the samples are large for the surveys of test-takers and the public, with from about 600 to 2,300 test-takers and from about 1,000 to more than 1,700 members of the public. But the samples are relatively small for admissions officers, ranging from roughly 100 to 200, and for breakdowns by gender, ethnicity, and income of test-takers and the public. In the tables that follow, I have drastically abbreviated the questions and shown the year that the survey was actually done, not the year it may have been published.

I begin with test-takers' attitudes. Table 1 shows the findings for five surveys, conducted between 1978 and 1997–1998, of test-takers' attitudes about the validity of admissions tests in general and of the SAT[®] and the ACT in particular. A large majority with positive attitudes in the late 1970s dropped to a bare majority by the late 1990s.

Table 1

Test Takers' Attitudes About Validity of College Admissions Tests

Year	Question (abridged)	Percent agreeing
1978 ^a	The SAT-Verbal was fair as a test of your abilities.	73
	The SAT-Math was fair as a test of your abilities.	81
1981–1982 ^b	The SAT-Verbal was fair as a test of your abilities.	71
	The SAT-Math was fair as a test of your abilities.	75
1986 ^c	College admissions tests predict freshman-year grades.	61
	The SAT was a fair measure of your academic ability.	62
	The ACT was a fair measure of your academic ability.	71
1996 ^d	The SAT measured your ability to perform in college.	51
1997–1998 ^e	The SAT measured your ability to perform in college.	54

^aResponse Analysis Corporation (1978). ^bBaird (1984). ^cGallup Organization (1987). ^dGordon S. Black Corporation (1996). ^eGordon S. Black Corporation (1998).

Only two surveys, from 1981–1982 and 1986, provided breakdowns for subgroups (see Table 2). Men and women, and high- and low-income test-takers differed minimally in their attitudes about validity. White and Hispanic test-takers also differed minimally, except for a slight difference for the ACT—fewer Hispanic test-takers had positive attitudes. But Black test-takers' attitudes differed slightly from the other ethnic groups—fewer Black test-takers had positive attitudes (in one instance, slightly less than half.).

The findings for three surveys, conducted between 1986 and 1997–1998, of test-takers' attitudes about unbiasedness in admissions tests in general or specifically in the SAT and ACT appear in Table 3. Attitudes varied considerably, ranging from a slight majority to a large majority with positive attitudes. However, the 96% figure in the 1997–1998 survey certainly looks like an outlier.

Table 2***Test Takers' Attitudes About Validity of College Admissions Tests: Subgroup Differences***

Question (abridged)	Gender		Ethnicity			Income	
	Men	Women	White	Black	Hispanic	Low	High
	1981–1982 ^a						
The SAT-Verbal was fair as a test of your abilities.						70	70
The SAT-Math was fair as a test of your abilities.						76	72
	1986 ^b						
College admissions tests predict freshman grades.	61	60	62	52	62	61	59
The SAT was a fair measure of your academic ability.	67	58	64	44	60	62	65
The ACT was a fair measure of your academic ability.	76	67	74	56	64	71	74

Note. Percent agreeing is shown.

^aBaird (1984). ^bGallup Organization (1987).

Table 3***Test Takers' Attitudes About Bias in College Admissions Tests***

Year	Question (abridged)	Percent agreeing
1986 ^a	College admissions tests are fair, not biased.	66
	The SAT is fair, not biased against minorities.	63
	The ACT is fair, not biased against minorities.	67
1996 ^b	The SAT was appropriate for all people, regardless of sex and ethnicity.	88
1997–1998 ^c	The SAT was appropriate for all people, regardless of their ethnicity.	96

^aGallup Organization (1987). ^bGordon S. Black Organization (1996). ^cGordon S. Black Corporation (1998).

Only one survey, from 1986, provided subgroup breakdowns (see Table 4). Gender differences were minimal, apart from a slight difference for the SAT—fewer women had positive attitudes. Income differences were also minimal, except for a slight difference for admissions tests in general—more high-income test-takers had positive attitudes. White and Hispanic test-

takers also differed minimally, but Black test-takers differed slightly or moderately from the other ethnic groups—fewer Black test-takers had positive attitudes (in two cases, slightly fewer than half). This is the same pattern of ethnic group differences observed for attitudes about validity.

Table 4

Test Takers' Attitudes About Bias in College Admissions Tests: Subgroup Differences

Questions (abridged)	Gender		Ethnicity			Income	
	Men	Women	White	Black	Hispanic	Low	High
	1986 ^a						
College admissions tests are fair, not biased.	70	63	69	40	72	55	65
The SAT is fair, not biased against minorities.	68	57	66	40	66	64	62
The ACT is fair, not biased against minorities.	69	66	70	59	66	65	70

Note. Percent agreeing is shown.

^aGallup Organization (1987).

What about the public's attitudes? Table 5 reports the results for two surveys in 1986 and 1997 of attitudes about the validity of admissions tests in general. In both surveys, a bare majority had positive attitudes. No subgroup breakdowns are available.

Table 5

The Public's Attitudes About Validity of College Admissions Tests

Year	Question (abridged)	Percent agreeing
1986 ^a	College admissions tests are a fair measure of academic ability needed for success in college.	54
1997 ^b	College admissions tests predict success in college.	52

^aOpinion Research Corporation (1997). ^bRowan & Blewitt Incorporated (1998).

The findings for three surveys, from 1986 to 1997, of the public’s attitudes about unbiasedness of admissions test in general appear in Table 6. Between a bare and a modest majority had positive attitudes.

Table 6

The Public’s Attitudes About Bias in College Admissions Tests

Year	Question (abridged)	Percent agreeing
1986 ^a	College admissions tests are fair, not biased against minorities.	68
1995 ^b	College admissions tests are an unbiased measure of all applicants, not biased against minorities.	53
1997 ^c	It is fair to use the same college admissions test to determine the ability of all applicants, including minorities.	70

^aGallup Organization (1986) ^bYankelovich Partners (1995). ^cRowan & Blewitt Incorporated (1998).

Breakdowns by ethnicity are available for surveys from 1986 and 1997 (see Table 7). The differences among White, Black, and Hispanic test-takers were minimal in the 1986 survey, but White and Black test-takers differed moderately in the 1997 survey—fewer Black test-takers had positive attitudes (well under half). Data for Hispanic test-takers are unavailable for this survey.

Table 7

The Public’s Attitudes About Bias in College Admissions Tests: Ethnic Differences

Questions (abridged)	Subgroup		
	White	Black	Hispanic
	1986 ^a		
College admissions tests are fair, not biased against minorities.	69	60	65
	1997 ^b		
College admissions tests are an unbiased measure of all applicants, not biased against minorities.	53	28	

Note. Percent agreeing is shown.

^aGallup Organization (1986). ^bYankelovich Partners (1997).

Finally, I come to college admissions officers. Surprisingly, little data are available about the attitudes of these critical stakeholders. Table 8 shows the results for the two available surveys, from 1996 and 1997, of college admissions officers' attitudes about the validity of admissions tests in general. The surveys, just a year apart, differ dramatically: A large majority in the first survey but only a bare majority in the second had positive attitudes.

The findings on their attitudes about unbiasedness of admissions tests in general are also given in Table 8 for the only available survey, from 1997. A slight majority had positive attitudes.

Table 8

Admissions Officers' Attitudes About Validity and Bias of College Admissions Tests

Year	Question (abridged)	Percent agreeing
Validity		
1996 ^a	The SAT differentiates students for admissions decisions.	89
1997 ^b	College admissions tests predict success in college.	51
Bias		
1997 ^b	It is fair to use the same college admissions test to determine the ability of all applicants, including minorities.	58

^aGordon S. Black Corporation (1996). ^bRowan & Blewitt Incorporated (1998).

It is obvious that the data I have hurriedly reviewed give no more than an incomplete and perhaps superficial picture of attitudes about college admissions tests held by the different constituencies. Attitudes may have changed, for the better or worse, since many of the surveys were done in the 1980s and 1990s. (The most recent survey was done in 1997–1998.) And the results based on wildly different questions cannot be compared with much confidence. All in all, though, it appears that acceptance of college admissions tests is no more than moderate across the board. I have not scrutinized the surveys of admissions tests used in graduate and professional schools, but my impression is that those results are similar.

Less than perfect acceptance of admissions tests is not surprising; indeed, it is probably inevitable. People have unrealistic views about academic ability and academic intelligence. Baird (1977) found that test-takers in his survey complained that important “abilities” were not tapped by the admissions tests that they had taken for graduate and professional school. The abilities that they thought were missing included creativity, academic motivation, and interpersonal skills.

Similarly, Sternberg and his coworkers (Sternberg, Conway, Ketron, & Bernstein, 1981) reported that the general public saw academic intelligence as being made up of verbal ability, practical problem-solving ability, and social competence. When people define academic ability so expansively, they can only think that admissions tests are inadequate.

Furthermore, test-takers, in particular, rightly see taking a test as an occasion for being judged, as Fiske (1967) pointed out some years ago. Test-takers are being evaluated and may be found wanting, as he put it. It is natural that they resent being put in such an aversive situation.

Test-takers' attitudes may also be affected by a *self-serving bias*: they may attribute their poor performance to invalidity in the test, as Chan and his colleagues (Chan, Schmitt, Jennings, Clause, & Delbridge, 1998) noted. Indeed, the 1981–1982 survey of SAT test-takers' attitudes about validity supports this point: The low scorers had slightly less positive attitudes.

The less positive or even distinctly negative attitudes of Black test-takers and Black members of the public, particularly about whether the tests are unbiased, are of great interest and importance. Of course a self-serving bias may play some role, given the lower test scores, on average, for Black test-takers. What else contributes to these negative attitudes needs to be addressed.

None of these findings should come as a great surprise, for the same things have been observed in laboratory studies (e.g., Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Gilliland, 1994). But it is reassuring and very persuasive that the results from small convenience samples of college sophomores and the like are replicable in large, representative, national samples of test-takers and the public.

In line with the recommendations by Brim and his coworkers (Brim, Glass, Neulinger, & Firestone, 1969) four decades ago and by Baird (1977) three decades ago, more needs to be done to educate all the constituencies in the testing enterprise, test-takers as well as academic institutions and the public, about what admissions tests are able—and unable—to do. Such efforts would enhance test-takers' motivation and improve the tests' acceptance by all parties.

These stake-holders' attitudes about admissions tests need to be assessed on a continuing basis, as Ryan and Ployhart (2000) proposed that industry do with employment tests. Although ETS commissioned several surveys in the 1980s and 1990s, not one of the publishers of admissions tests has a regular program for assessing attitudes. I am also astonished that none of the surveys that I have just described were done in this century. It would be a simple matter to

insert a few questions about admissions tests into the omnibus surveys that survey organizations do all the time. Using exactly the same questions would permit the tracking of changes over time.

A final methodological observation: This project illustrates the value of existing archives of survey data. This is not news to survey people or sociologists, but it may be to many in the testing community and to lots of other psychologists. In fact, this project does not do justice to the archives. The ready availability of the raw data makes possible analytical studies, not just the descriptive results cataloged here.

Stereotype Threat

The second project I want to talk about has to do with stereotype threat. Stereotype threat is a concern about fulfilling a negative stereotype regarding the ability of one's group when placed in a situation where this ability is being evaluated, such as when taking a cognitive test. These pervasive stereotypes target not just minorities and women, but also the working class, the elderly, and other groups. The result of this concern is that performance on the ability assessment is adversely affected (see Steele, 1997).

Research on stereotype threat began with Steele and Aronson's experiments in 1995 (Steele & Aronson, 1995). This phenomenon is illustrated by a classic study. Spencer, Steele, and Quinn (1999, Study 3) administered a test made up of quantitative ability items from the Graduate Management Admission Test to undergraduate men and women (see Figure 1). Stereotype threat was manipulated by telling one group that there was no gender difference on the test (the no-difference condition) and telling the other group nothing (the control condition). Women's scores were higher in the no-difference condition than in the control condition, while men's scores were unaffected.

Similar effects have been amply documented in investigations using a variety of manipulations of stereotype threat, a wide range of populations of research participants, and a number of different ability and achievement tests. Well over 200 studies have been published, almost all laboratory experiments.

This body of work suggests that stereotype threat is potentially a serious influence on test performance and a source of invalid variance in test scores. Just as important, it implies that stereotype threat may explain the deficits on standardized ability and achievement tests that are commonly found for minority groups, women (on quantitative tests), and working-class people.

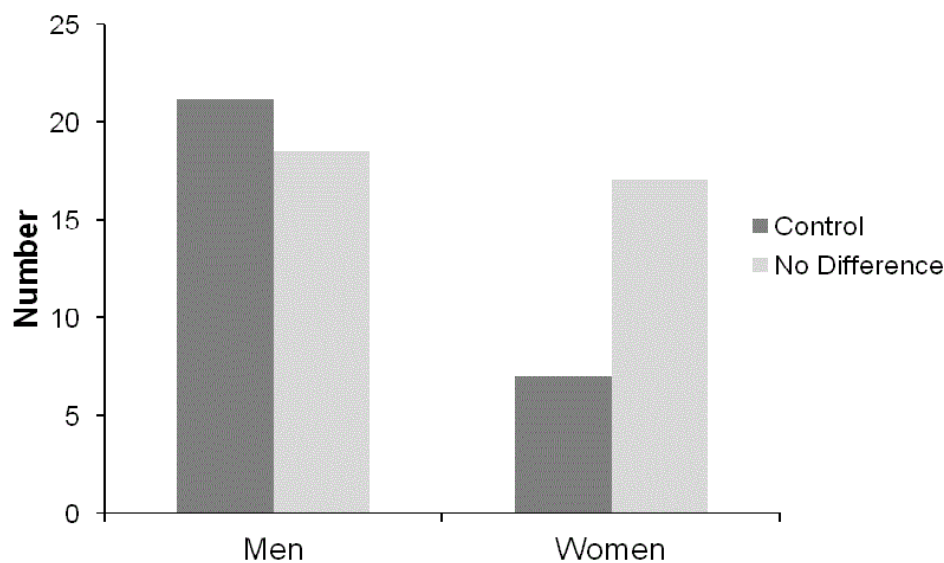


Figure 1. Mathematical score (items from GMAT). Data from Spencer et al. (1999, Study 3).

However, the jury is still out about the impact of stereotype threat on test performance outside of the laboratory. It is exceedingly difficult to study stereotype threat in the real world. As Cullen, Hardison, and Sackett (2004) pointed out, pragmatic and ethical concerns preclude the powerful manipulations, often produced by deceptions, that can be carried out in the laboratory.

A handful of studies with operational tests in high-stakes settings have failed to find evidence of stereotype threat—with one exception. The studies with negative findings are the field experiments that I did with Ward (Stricker & Ward, 2004) on an Advanced Placement[®] test and on the Computerized Placement Tests, and the predictive validity studies done by Cullen and his colleagues (Cullen et al., 2004; Cullen, Waters, & Sackett, 2006) with the SAT[®] and Armed Services Vocational Aptitude Battery (ASVAB) and real-life criteria: freshman GPA for college students and military job criteria for soldiers. The study with positive findings was a field experiment by Good, Aronson, and Inzlicht (2003) with a statewide test of minimum competency, the Texas Assessment of Academic Skills, which is used to determine whether students should be promoted.

Barack Obama’s presidential campaign provided a natural experiment for assessing stereotype threat outside of the laboratory. Laboratory studies have established that exposure to a

successful role model can cancel the effects of stereotype threat. For example, Marx and Roman (2002) used men and women experimenters portrayed as highly competent in mathematics to administer a test made up of quantitative ability items from the *GRE*[®] General Test to undergraduate men and women (see Figure 2). Women’s scores were higher when the test administrator was a woman than when the administrator was a man, while men’s scores were unaffected.

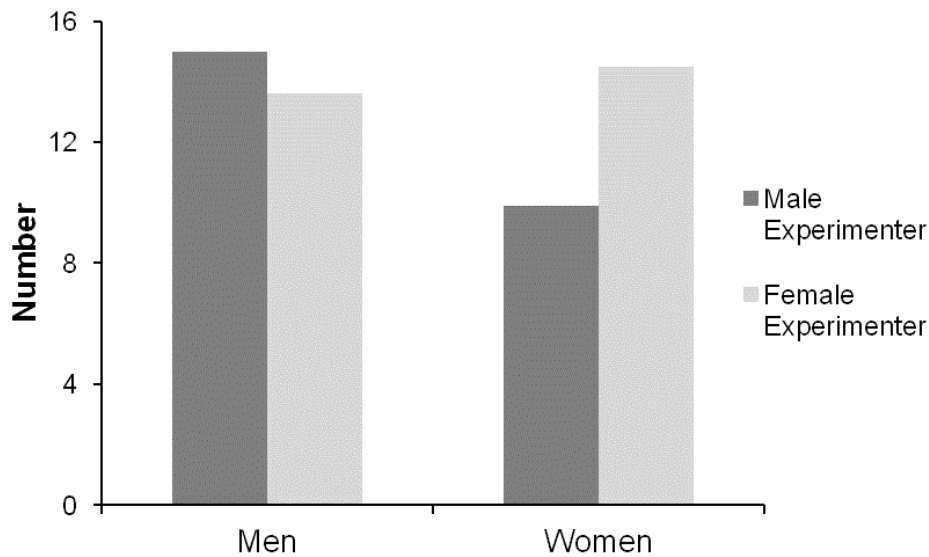


Figure 2. Mathematical score (items from GRE General Test)—covariance adjusted for SAT scores. Data from Marx and Roman (2002).

Marx, Ko, and Friedman (2009) set out to test this role model effect in a field study during the 2008 presidential campaign. They administered a test composed of verbal ability items from the GRE General Test to samples of Black and White adults drawn from a nationwide consumer behavior panel that participates in web-based surveys and experiments. Four samples were drawn at different points in 2008: in August, just *before* the Democratic convention; in August again, just *after* the convention; in October, a relatively uneventful period; and in November, just after the election. In addition, the after-convention sample in August was subdivided into those who watched Obama’s acceptance speech and those who did not see it.

The Black and White means on the test are shown in Figure 3. (The Black and White participants were matched on age and English proficiency, and education was covaried out of the

test scores.) The mean differences between the ethnic groups were statistically significant before the convention, after the convention for those who did not see the speech, and again in October. All of these mean differences were large, at least three-quarters of a standard deviation. But, remarkably, the mean differences were *not* significant after the convention for those who watched the speech or after the election. These mean differences were about a third of a standard deviation or less.

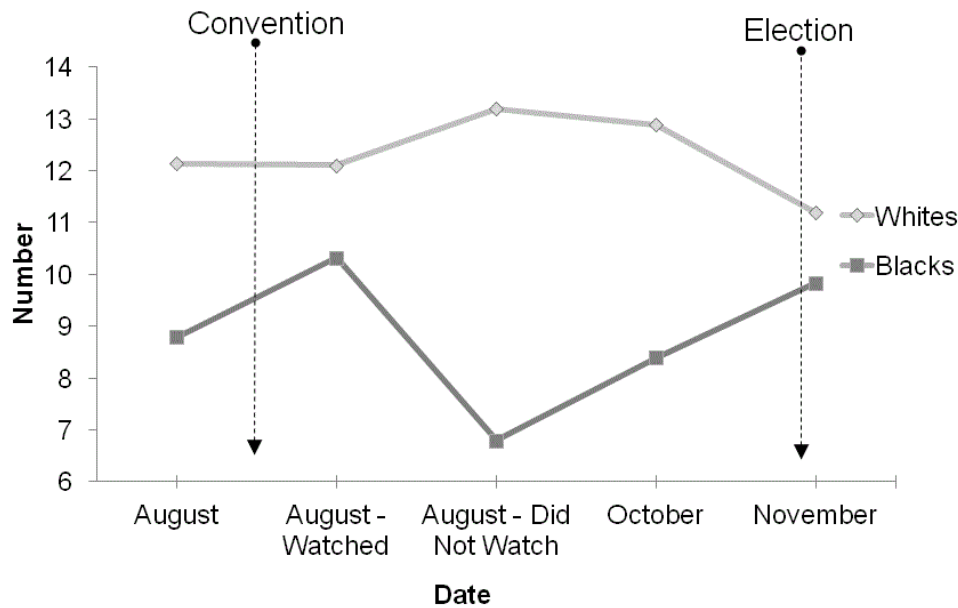


Figure 3. Verbal score (items from GRE General Test)—covariance adjusted for educational level. Data from Marx et al. (2009).

Marx and his colleagues (2009) speculated that the absence of a mean difference at these two points is due to the Black participants perceiving Obama as successfully achieving something concrete—winning the nomination and winning the election. And they argued that the typical mean differences between the ethnic groups persisted at the other points because Obama’s achievements were too general then and not a clear sign of him being a successful role model. In short, the mere salience of a role model is not enough to produce an effect.

These results, labeled the *Obama effect*, received a great deal of attention in the media and among at least some psychologists. It was seen as a dramatic demonstration of stereotype threat outside of the laboratory and of the power of Obama’s success in the campaign in buffering the effects of this threat and even shrinking the Black-White test-score gap. However,

the explanation for finding a difference at some points and not at others seems counterintuitive and is not supported by other research (Aronson, Jannone, McGlone, & Johnson-Campbell, 2009). And the sample sizes were extremely small. The biggest sample, the one before the convention, had 30 Black and 101 White participants. The smallest sample, those who did not watch the speech, had only 8 Black and 56 White participants.

When this study was first reported in *The New York Times* (Dillon, 2009)—it was subsequently published in the *Journal of Experimental Social Psychology*—a “testing expert” was quoted as saying, “It’s a nice piece of work. I’d want to see another study replicating their results before I get too excited about it” (p. A15). I could not help laughing when I read this comment: It might be a long time before the next minority politician comes out of nowhere and wins the presidency! But then it occurred to me that test publishers have troves of archival data, spanning the campaign period, that could be mined. I was fortunate to gain access to GRE General Test data for this study. (I am grateful to the GRE people for their responsiveness.) As you probably know, this test is administered daily, by computer, in testing centers around the world.

I obtained data for all Black and White test-takers (these are American citizens) in eight time periods during the campaign, starting in December 2007 and ending in November 2008. The time periods include the same four used by Marx and his coworkers (Marx et al., 2009), and four other relevant periods (see Figure 4). (Of course it is impossible to distinguish, as the Marx team did, between those who saw Obama’s acceptance speech and those who did not.) The new time periods were December 2007, before the Obama campaign took off after he won the Iowa caucus in January; February 2008, after Super Tuesday, when Obama won more states and delegates than Clinton; April, a relatively uneventful period without concrete successes by Obama, like the October period used by the Marx team; and June, after Clinton conceded.

In addition, I obtained data for the same eight periods in the previous 12 months, December 2006 to November 2007. I did this to control for natural changes in the test-taking cohorts, both within and between years. It is well established that the performance level on the GRE General Test varies considerably—and consistently—over the year. Year-to-year changes may also occur. Hence, different cohorts, say those tested in December 2007 and February 2008, are not random samples from the same population. All of the samples in both years were relatively large, with over 300 Black and 1,800 White test-takers.

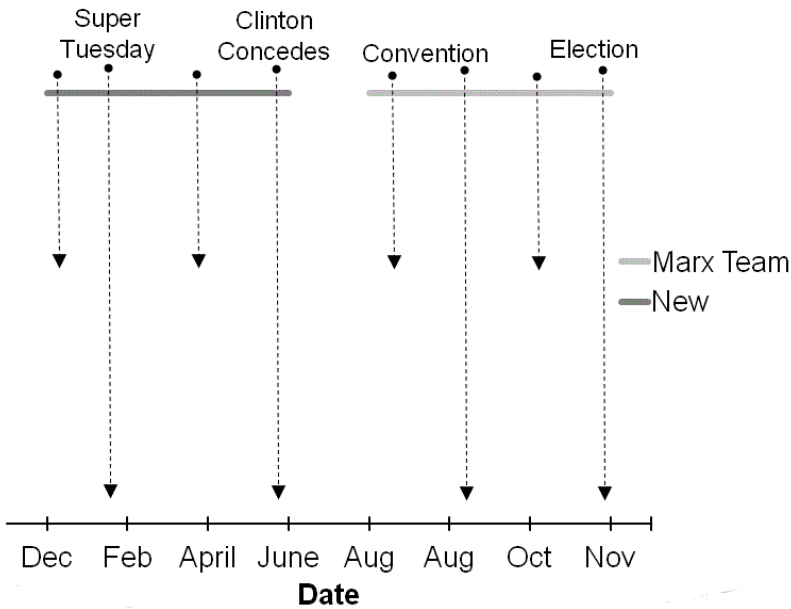


Figure 4. 2007–2008 data points for Marx et al. (2009) and new study.

I will describe only the results for the total sample of men and women and the Verbal Reasoning section—this is the parent of the ad hoc test used by the Marx team (Marx et al., 2009). (The results are similar for each gender and for the other sections of the test, Quantitative Reasoning and Analytical Writing.) The Verbal Reasoning section means for Black and White test-takers in 2007–2008 (and the .95 confidence intervals) are shown in Figure 5. It is apparent that the Black and White means were substantially different in all of the time periods. The differences ranged from 84 to 103 points on the GRE 200–800 score scale (the standard deviation was 111 on the Verbal Reasoning section for American citizens taking the test in 2006–2007; Educational Testing Service, 2008).

The means for 2006–2007 along with the means for 2007–2008 appear in Figure 6. The corresponding means for the two years were very similar, but in the second half of 2007–2008 two of the Black means were slightly higher, and one of the White means was slightly lower. (The largest difference was 19 points, about a fifth of a standard deviation.)

In short, the dramatic Obama effect observed by Marx and his colleagues (Marx et al., 2009) is not apparent in these data. The Black-White disparity in mean test scores persisted throughout the campaign and was similar to the mean differences the year before the campaign began. Naturally, a raft of differences exists between my study and the investigation by the Marx team that may explain the discrepant results. Two explanations are most likely. First and most

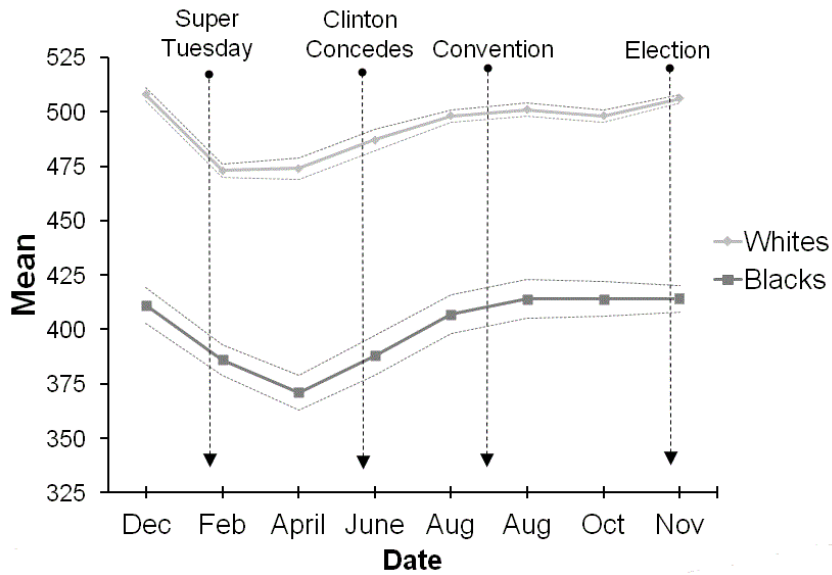


Figure 5. GRE Verbal Reasoning score, 2007–2008 (dashed lines depict .95 confidence limits).

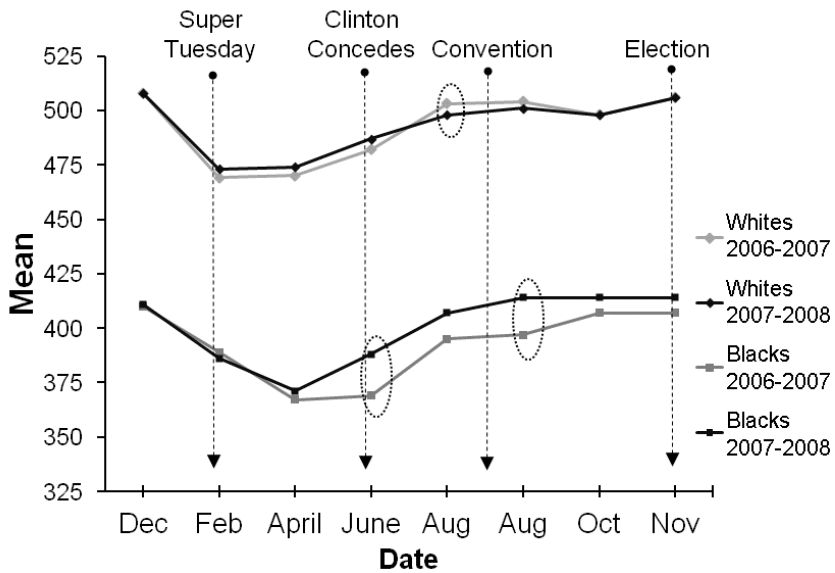


Figure 6. GRE Verbal Reasoning score, 2006–2007 and 2007–2008 (circles identify significant, $p < .05$, mean differences).

obvious, their study had very small samples, which may produce unreliable results. Second, it used research participants in a low-stakes research setting, whereas mine used applicants to graduate school taking a high-stakes admissions test. Motivation to perform well on the admissions test may have overridden any harmful effects of stereotype threat. This same explanation has been offered before to account for the negative results usually found in field studies of operational tests and the positive results consistently found in the laboratory (Stricker, 2008).

This study is not the only one that failed to support the findings by the Marx group (Marx et al., 2009). Aronson and his colleagues (Aronson et al., 2009) conducted an experiment in June and July 2008, after Clinton conceded but before the convention. Black and White undergraduates were prompted to think about Obama, McCain, or simply “an American politician” before they took a verbal section of the Medical College Admission Test. This manipulation did not significantly affect the test performance of either the Black or White students. Hence, the Obama effect was not evident.

My study joins previous investigations that failed to find a stereotype-threat effect on operational tests in high-stakes settings. Still, it would be a capital mistake to rule out stereotype threat as a baleful influence on such tests, given the difficulty of investigating this phenomenon in real life and the weight of findings from the laboratory. It would be prudent for the testing community to search for or create other opportunities to study stereotype threat on operational tests, as in the Cullen group’s correlational studies (Cullen et al., 2004, 2006), the field experiments by Ward and me (Stricker & Ward, 2004) and by Good and her colleagues (2003), and in the natural experiment provided by the Obama candidacy. I continue to be struck—and disheartened—by the ferment that stereotype threat has produced among social psychologists and the public, in sharp contrast to the relative indifference among testers.

Conclusions

Two themes run through these two projects. One is that social and intrapersonal influences are important phenomena, but are largely neglected by my testing brethren. The other theme is that it is both feasible and valuable to use real-life data to study them. Survey organizations and test publishers collect enormous amounts of high-quality data, and their archives are extraordinary but woefully underutilized resources. The data in these repositories

are grist for many worthwhile investigations. They include what might be thought of as historical psychology that links notable events with attitudes and test performance, as was done in the Obama effect studies.

A modest shift in the testing community's priorities from psychometrics—and statistics—to the social and intrapersonal side of testing might pay real dividends, not only enhancing the validity of tests but also improving the viability of the testing enterprise itself. I hope that we will have a chance to see this hypothesis tested.

References

- Aronson, J., Jannone, S., McGlone, M., & Johnson-Campbell, T. (2009). The Obama effect: An experimental test. *Journal of Experimental Social Psychology, 45*, 957–960.
- Baird, L. L. (1977). What graduate and professional school students think about admissions testing. *Measurement in Education, 7*(3), 1–7.
- Baird, L. L. (1984). Relationships between ability, college attendance, and family income. *Research in Higher Education, 21*, 373–395.
- Brim, O. G., Jr., Glass, D. C., Neulinger, J., & Firestone, I. J. (1969). *American beliefs and attitudes about intelligence*. New York, NY: Russell Sage Foundation.
- Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*, 300–310.
- Chan, D., Schmitt, N., Jennings, D., Clause, C. S., & Delbridge, K. (1998). Applicant perceptions of test fairness: Integrating justice and self-serving bias perspectives. *International Journal of Selection and Assessment, 6*, 232–239.
- Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theories. *Journal of Applied Psychology, 89*, 220–230.
- Cullen, M. J., Waters, S. D., & Sackett, P. R. (2006). Testing stereotype threat theory predictions for math-identified and non-math-identified students by gender. *Human Performance, 19*, 421–440.
- Dillon, S. (2009, January 23). Study sees an Obama effect as lifting Black test-takers. *The New York Times*, p. A15.
- Educational Testing Service. (2008). *Factors that can influence performance on the GRE General Test, 2006-2007*. Princeton, NJ: Author.
- Fiske, D. W. (1967). The subject reacts to tests. *American Psychologist, 22*, 287–296.
- Gallup Organization. (1986). *Report on national survey of ETS awareness for Educational Testing Service*. Princeton, NJ: Author.
- Gallup Organization. (1987). *The Gallup national survey of recent college admissions test-takers*. Princeton, NJ: Author.

- Gilliland, S. W. (1994). Effects of procedural and distributive justice on reactions to a selection system. *Journal of Applied Psychology, 79*, 691–701.
- Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Applied Developmental Psychology, 24*, 645–662.
- Gordon S. Black Corporation. (1996). *Educational Testing Service 1996 customer satisfaction benchmarking project*. Rochester, NY: Author.
- Gordon S. Black Corporation. (1998). *Educational Testing Service 1997-98 customer satisfaction measurement*. Rochester, NY: Author.
- Herriott, P. (1989). Selection as a social process. In M. Smith & I. Robertson (Eds.), *Advances in selection and assessment* (pp. 171–187). Chichester, England: Wiley.
- Iles, P. A., & Robertson, I. T. (1989). The impact of personnel selection procedures on candidates. In P. Herriott (Ed.), *Assessment and selection in organizations: Methods and practice for recruitment and appraisal* (pp. 257–271). Chichester, England: Wiley.
- Lerner, B. (1980). The war on testing: David, Goliath & Gallup. *Public Interest, 60*, 119–147.
- Marx, D. M., Ko, S. J., & Friedman, R. A. (2009). The “Obama Effect”: How a salient role model reduces race-based performance differences. *Journal of Experimental Social Psychology, 45*, 953–956.
- Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin, 28*, 1183–1193.
- Nevo, B. (1993). The practical and theoretical value of Examinee Feedback Questionnaire (EFeQ). In B. Nevo & R. S. Jager (Eds.), *Educational and psychological testing: The test-taker's outlook* (pp. 85–113). Toronto, Canada: Hogrefe & Huber.
- Opinion Research Corporation. (1997). *American attitudes toward higher education, 1986: Results of a comprehensive nationwide survey*. Princeton, NJ: Author.
- Phelps, R. P. (1998). The demand for standardized student testing. *Educational Measurement: Issues and Practice, 17*(3), 5–23.
- Phelps, R. P. (2003). *Kill the messenger: The war on standardized testing*. New Brunswick, NJ: Transaction.
- Phelps, R. P. (2005). Persistently positive: Forty years of public opinion on standardized testing. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 1–22). Mahwah, NJ: Erlbaum.

- Response Analysis Corporation. (1978). *SAT monitor program: High school students view the SAT and college admission process*. Princeton, NJ: Author.
- Rowan & Blewitt Incorporated. (1998). *Educational Testing Service nationwide opinion survey final report*. Washington, DC: Author.
- Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management*, 26, 565–606.
- Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35, 4–28.
- Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52, 613–629.
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797–811.
- Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology*, 41, 37–55.
- Stricker, L. J. (2008). *The challenge of stereotype threat for the testing community* (ETS Research Memorandum No. RM-08-12). Princeton, NJ: Educational Testing Service.
- Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test-takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology*, 34, 665–693.
- Yankelovich Partners. (1995). *Time/CNN/Yankelovich Partners poll, March 1995*. [Data file, USYANKP.95007.Q26]. Retrieved from <http://www.ropercenter.uconn.edu>
- Yankelovich Partners. (1997). *Time/CNN/Yankelovich Partners poll, September 1997*. [Data file, USYANKP.111797.R30]. Retrieved from <http://www.ropercenter.uconn.edu>
- Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. New York, NY: Routledge Falmer.