EDITED BY

TUKUFU ZUBERI AND EDUARDO BONILLA-SILVA

WHITE LOGIC, WHITE METHODS

RACISM AND METHODOLOGY

5

Causation and Race

Paul W. Holland

For over 2000 years, ideas about causation have been discussed, classified, and criticized. To mention only a few of the most influential authors: in philosophy there are Aristotle, Hume, and Mill; in medicine there are Koch and Henle (Yerushalmy and Palmer 1959) and Sir A. Bradford Hill (1965); and in social science research and program evaluation there are Campbell and Stanley (1963). With all this work explaining, refining, and clarifying what causation means and how to distinguish it from "mere association," it is still worth repeating the maxim: before one leaps to a causal conclusion, one needs to consider first the other noncausal explanations and eliminate them.

Two of the most commonly occurring alternatives to causal explanations are reverse causation and common causes (or hidden confounding). Following are two examples that are easy to identify, yet continue to cause problems in educational policy discussions.

Example 1 (Reverse Causation): It is easy to find data, for example from the National Assessment of Educational Progress (NAEP), in which widely accepted educational materials and practices (such as dividing classes into reading groups, using work sheets, and repetitive drill and practice) are associated with lower student performance on NAEP. The causal explanation is that these practices inhibit student learning and need to be replaced in school reform efforts. The noncausal explanation is that those who need more help may get it in a caring and student-oriented system of instruction and their low test performance is only indicative of their need, and not of

the result of the practices. The noncausal explanation is "reverse causation" in the sense that the apparent effect, that is, low test scores, is actually a measure of the cause, that the students are being taught in particular ways, rather than being the real effect of these practices.

Example 2 (Common Causes): It is equally easy to find NAEP data that shows that socially desirable educational practices (such as smaller classes, computers, and low teacher turnover) are associated with higher student test performance. The causal explanation is that these desirable things are desired because they are good for students and help them to succeed academically. The noncausal alternative explanation is that what we are really seeing is social segregation and socioeconomic status (SES) differences, which, like it or not, are associated with (might even cause!) both higher scores and more socially desirable schooling conditions. The common cause is SES differences. In this example it is possible that the educational practices are making a positive difference in the education of the students, but until the effect of SES is sorted out, the amount of this difference is difficult to know.

These examples of the problems with causal explanations are only two from a long list, but they are easy to identify and understand. Both are special cases of Simpson's paradox, which has perplexed users of statistics for over 100 years (Simpson 1951; Pearl 2000). Simpson's paradox says that a correlation or association between two variables can change in quite dramatic ways when the effect of a third variable is taken into consideration. A famous U.S. example is the claim by the UC Berkeley student newspaper that graduate student admissions at Berkeley were biased against women. The data showed exactly that. The Berkeley-wide rate of acceptance of women graduate students was lower than that of men. However, when the departments to which the students were applying were examined it was discovered that men and women applied to different departments and, interestingly, the different graduate programs admitted students at different rates. Women tended to apply to the departments where the acceptance rates were lower. In fact at the department level, there was a slight tendency to admit women at a higher rate than men (Bickel, Hammel, and O'Connell 1975). The third variable here was "department applied to," and a third variable, associated with the two of interest (gender and admission), can do amazing things with the original association.

I think one way to understand Mark Twain's attribution to Queen Victoria's prime minister, Benjamin Disraeli, of that famous put-down, "Sir! There are three kinds of lies—lies, damn lies and Statistics!" is as the blustering reaction of a great politician to someone's (mis)use of statistics to trash one of his pet policies. Simpson's paradox no doubt abounded in the trade and currency data that Disraeli and others needed for policy analysis

in the middle of the nineteenth century. Who knows what the "third variable" was or the data presented or Disraeli's pet policy (Twain never tells us), but rest assured, no matter how gifted an orator he was, Disraeli was doomed to naught but uttering pure bluster if it was Simpson's paradox he was up against. British statisticians Pearson and Yule were the first to understand the workings of Simpson's paradox but only did so years after Disraeli had left office.

However, the topic of this chapter is not about the type of misplaced causal thinking that reverse causation, common causes, and Simpson's Paradox exemplify. My interest here is one that has concerned me for some time (Holland 1988b) and that is of an order different from the simple misidentification of association for causation. The problem I wish to address is: What is the causal role of variables such as "race" in social science research?

Every day, an economist, a sociologist, or a political scientist "runs" a regression analysis in which some variable denoting the race of the person who is the unit of analysis appears as a predictor (along with other predictors) of some outcome variable. Every day, the analyst interprets the coefficient of this race variable as the "effect of race" on the outcome variable. Is there a "causal interpretation" to this race effect?

My answer to this question is that race is not a causal variable and for this reason "race effects" per se do not have any direct "causal" interpretation. It is also clear, however, that race does play an important role in some causal studies and that more clarity as to what this role is will help us understand concepts like "discrimination" and "bias" in ways that make fruitful use of causal ideas. In the rest of this chapter I will give the details of my argument and point of view, and illustrate it with an example of biased tests in a later section.

One warning. Those who wish a serious discussion of the meaning of race will have to look elsewhere. I take racial categories, however determined, as given. This is also the plight of the analyst who runs his or her regressions. For the most part, someone else determines the definition of the race variable and the analyst has to use the available data. I do not apologize for this superficiality on my part, because it is the common superficiality of those who employ race as a variable in their analysis. While not satisfactory for every situation, this approach to race is good enough for many purposes, or at least, the alternatives are even less satisfactory.

For the record, I regard race as a socially determined construction with complex biological associations. I also believe that it is very naive to disregard the durability and power of social constructions. Be that as it may, race is not a neutral concept and its many consequences for social interaction and other activities are the subject of a vast literature to which this chapter will not contribute.

CAUSATION

In this section I will give a relatively brief discussion of a few essential points about causation that are germane to my point of view. Related discussions are in Holland (1986, 1988a, 1988b, 2001). To begin, it is useful to distinguish between two classes of scientific studies in the social sciences: descriptive and causal studies.

Descriptive Studies

Descriptive studies have the goal of describing some phenomenon or state of affairs. Typical examples are ethnographic studies of a social system, detailed classroom observations, or sample surveys of characteristics of a population. The most ubiquitous type of purely descriptive study in American life is the opinion poll. Polls have the sole purpose of describing current opinion/sentiment of some population on some set of relevant issues. In education research the most important current descriptive studies are the National and State surveys collectively called the National Assessment of Educational Progress (NAEP). An early important education survey was the "Coleman Report" (Coleman 1966) and there are also important longitudinal surveys as well, such as High School and Beyond and the various versions of The National Educational Longitudinal Study (NELS).

It sometimes helps to classify questions in terms of the interrogatives of English. Those most relevant to descriptive studies are "Who? What? Where? and When?" The output of a descriptive study is a description, be it a "thick description" of some event or phenomenon or merely a mean, a distribution, or a correlation. An important contribution of statistics to descriptive studies is the twentieth-century invention of the sample survey employing random selection. "Careful observation and description" does, however, have ancient scientific credentials.

The Slippery Slope toward Causation

Description often results in comparisons, and a comparison often invites one to ask the other kinds of interrogatives, in particular, Why? or How? While these causal questions are, in some sense, more fundamental than those related to description, I remind myself regularly that, at least in the social sciences, casual comparisons inevitably initiate careless causal conclusions.

It is not unusual for our desire to know "Why?" to outstrip our ability to provide an adequate answer. For example, we may know that there are a variety of replicable differences in test performance between various groups of examinees (e.g., males and females or ethnic/racial groups), but why these

differences consistently arise often eludes serious explanation. In a related setting, NAEP's descriptive data are used time and again to address causal questions, and absurd conclusions can result from the failure to recognize the survey/descriptive nature of NAEP. The two examples, of reverse causation and common causes, given earlier are typical of this overenthusiasm for causal explanations.

Different types of research studies can make it more or less difficult to clearly distinguish between description and causation. We have somewhat pejorative language for this failure: "Correlation does not necessarily imply causation" and "mere correlational research." In my opinion, however, good descriptive studies, which lay out important dimensions of some social science phenomena, are highly underrated. On the other hand, there is a sense in which all studies are just descriptive studies and all that is ever observed in any study is "mere correlation." In this view some of these correlations have causal relevance while others do not. As my colleague Howard Wainer once quipped, "Where there is correlational smoking there may be causational cancer."

Causal Studies

I do not think that it is very useful to try to make an exhaustive catalog of all possible causal studies. Rather I think it is more helpful to try to recognize when a study has a "causal" focus, rather than being solely concerned with pure description of phenomena. Even studies that start out as purely descriptive can be given an apparent causal focus as we slide down the slippery slope initiated by casual comparisons. In such situations it is best to be wary of the slippage from description to causation.

Again, the interrogatives of English can begin to help us (though they have limitations). Questions of Why? or How? invite causal explanations. I believe, however, that there are really three distinct types of causal questions, with Why? and How? associated with only two of them. Confusing the three types of causal questions (or their answers) can make causal discussions confusing and contentious.

I call the answers to the three types of causal questions (a) Identifying Causes, (b) Assessing Effects, and (c) Describing Mechanisms. Let me amplify each of these in turn.

Identifying Causes. This is the usual answer to Why? A singular event occurs, and we seek its cause. "Why did the car (or stock market) crash?" "What caused his death?" "Why are test scores down?" There can be an element of blame in answers to Why-questions. For example, "Test scores are down because our curricula are a mile wide and an inch deep!" Legal responsibility can also be involved, as in assessing financial responsibility for an accident. Causal identification is often a form of speculative postmortem.

Assessing Effects. This is the answer to the missing type of causal question alluded to above. I think that "What if?" better describes the questions whose answers require the assessment of the effects of certain causes. Likewise, "What is?" is the proper form of the "What" questions that descriptive studies can address. When we ask a What-if question we seek to know the effect of some cause or intervention that we might contemplate making. "Will test scores go up if we reduce class size?" "What will happen to dropout rates if we end social promotion as we know it?" I think that the questions that are most relevant to the intersection of social science and public policy are these What-if questions whose answers involve assessing effects of causes or interventions.

Perhaps the simplest image that underlies our understanding of causal attribution is the comparison of two identical units of study, one exposed to one experience and the other exposed to another experience, which are then subsequently compared on an identical outcome criterion. Because these units of study are identical/similar to begin with and are evaluated in an identical/similar manner at the end, whatever difference is observed between them in the outcome is attributable to the differences they had in their intervening experiences and to nothing else.

An important contribution of statistics to the study of causation is the other twentieth-century invention of the randomized comparative experiment. Such study designs remove the need for finding "identical" units of study. They started in agriculture and quickly spread to many areas of science where uncontrollable variation in experimental material—the weather, the fertility of the earth, population, and so on—is a fact of scientific research life. In my opinion, such studies are very good at addressing What-if questions, but, on occasion, when coupled with predictions, they may indirectly address the speculations of Why? and How? as well.

There are deep formal connections between sample surveys that employ random selection of units for inclusion in the sample and comparative experiments that employ random assignment of units to experimental conditions. These two types of studies, however, address very different types of questions (the former descriptive, while the later causal) and ought not to be confused with each other.

Describing Mechanisms. This is the answer to How? We see that some effect follows from some cause and we want to know "How does it work?" "How does the effect arise from the action of the cause?" "How do smokerings form?" "How will class size reduction improve test scores?" "How does aspirin reduce heart attacks?" Understanding and identifying causal mechanisms is, perhaps, the primary driving force of science. Causal mechanisms are the closest things to "theory" that I will discuss here. Furthermore, causal mechanisms are almost always involved in that hallmark of science, prediction. "Describing causal mechanisms," like "identifying

causes," almost always involves some element of speculation—sometimes a very healthy dose of it.

The description of a causal mechanism can be completely wrong while at the same time the effect of the cause is clear and replicable. A well-known medical example concerns taking aspirin to reduce one's risk of heart attacks. The data on the reduction are clear and well established by a large randomized clinical trial. But at first the mechanism by which the reduction was achieved was in question. Was it aspirin's blood-thinning effect or its inflammation-reduction effect? Early explanations emphasized blood thinning, but later experimental work confirmed inflammation reduction. However tentative, causal mechanisms are often useful ways to encode our thinking about causal relationships (e.g., the germ theory of disease).

I think that it is important to be clear as to what type of questions a study is trying to or can answer: descriptive or causal; and, if causal, which type? One of the problems of communication between social scientists and policy makers is related to the distinction I make between assessing effects and describing mechanisms. Understanding some aspect of a causal mechanism often advances science (i.e., theory), whereas the needs of public policy often require an answer that assesses the effects of an intervention, rather than reasons or speculations as to how these effects come about. If class size reduction results in better student learning, a policy maker might argue that it does not matter if this effect is due to more time for individualized instruction, fewer classroom disruptions, or something else. On the other hand, the mechanism might matter to the policy maker if other reform policies besides class size reduction are of interest. Knowledge of the causal mechanism could indicate that other policies would be supportive or possibly contraindicated when classes are small. My view is that both positions need to be clearly delineated and not confused with each other.

Causal Variables

It should be clear, but it often is not, that the language of causation is more precise when we are concerned with assessing effects than when we are concerned with either identifying causes or with proposing causal mechanisms. In the latter two cases, "anything" can be a cause, because we are just talking rather than doing. When we design an experiment, however, the only things that can qualify as causes are "treatments" or "interventions." I think that putting limits on "what a cause can be" by using What-if questions to do this is useful and a very important step because it focuses on "doing" rather than on the (sometimes casual) causal talk of identifying causes and proposing causal mechanisms.

Long ago, Donald Rubin and I made up the slogan: No causation without manipulation (Holland 1986). Its purpose was to emphasize the ambiguity

that arises in causal discussions when things that were not treatments or interventions of some sort are elevated to the status of "causes." Not everyone agrees with this point of view (Marini and Singer 1988), but I still think it is a sound position and reiterate it here.

Our slogan closely corresponds to the following basic image, already mentioned, for understanding causation. Two identical/similar units of study, one exposed to one experience and the other exposed to another experience, which are then subsequently compared on an identical/similar outcome criterion. In this basic setting, the attribution of cause is to the different experiences (to which either unit could have been exposed), and not to some other characteristic of the units of study, because the units are "identical/similar." We can manipulate these experiences and thus attribute to them causation of any subsequent observed differences without necessarily suggesting a mechanism to explain the resulting difference. Thus, we return to my insistence that causes are experiences that units undergo and not attributes that they possess: No causation without manipulation!

Causal variables are those that reflect such manipulations or varying experiences between units of study. For a causal variable it is meaningful to ask about both (a) the result that obtained under the experience the unit was actually exposed to; and (b) the result that would have obtained had the unit been exposed to another experience. This is the essence of the definition of a causal effect. It inherently involves the use of counterfactual conditional statements (the result that would have obtained had the unit been exposed to another experience) (Lewis 1973). Properties or attributes of units are not the types of variables that lend themselves to plausible states of counterfactuality. For example, because I am a White person, it would be close to ridiculous to ask what would have happened to me had I been Black. Yet, that is what is often meant when race is interpreted as a causal variable.

There is no cut-and-dried rule for deciding which variables in a study are causal and which are not. In experiments, in which we actually have the control to manipulate conditions, there is usually no problem in identifying the causal variables (but even there, however, "what was actually manipulated" may not be so clear—perhaps the most famous examples being those involved with placebo effects).

Causal studies may also involve many types of nonexperimental settings in which we do not have control over which units are exposed to which experiences. In these cases it can become a challenge to determine what qualifies as a causal variable in the sense that I am using the term. The only rule I have is that if the variable could be a treatment in an experiment (even one that might be impossible to actually pull off due to ethical or practical issues) then the variable is probably a cause, and correctly called a causal variable. From this point of view, attributes of indi-

viduals such as test scores, age, gender, and race are not causes and their measurement does not constitute a causal variable.

Causation as a status symbol: We might ask why is it important to make a distinction between causal and noncausal variables? A biostatistician, whose name I have unfortunately misplaced, once made the telling point to me that in medical research it is highly valued to be able to assert that an association between one thing and another is "causal." However, he argued, as far as action is concerned it often does not matter whether the association is causal or noncausal. In medicine, "risk factor" refers to either case. Is high blood pressure causal in its association with heart disease, or are they both just due to a common cause? No matter, try to lower your blood pressure by diet, exercise, or drugs and you will probably be healthier. Being able to assert that the association is based on a causal connection is, in many circumstances, merely a status symbol, one that confers importance to the finding without any consequence for improved public health. Causes are sometimes easily related to action and noncauses are often not. For example, the physician can help you stop smoking, but not get younger!

From this point of view, which I believe is a healthy antidote to the search for a causal "Good Housekeeping Seal of Approval" on associations, it is the use of an association for important purposes that is its enduring value, and not its status as a causal variable.

IS RACE A CAUSE?

From the arguments of the last section, it should be clear that variables like race are not easily thought of as describing manipulations, and so, in my opinion, they do not qualify as causal variables. In this sense race is not a cause. It is important, however, to state the limitation of this assertion. Race is not a cause because race variables do not have causal effects as defined above. "What would your life have been had your race been different?" is so far from comprehensible that it is easily viewed as a ridiculous question. Few experimenters have manipulated race and, when they try to, it is a poor imitation of the real thing.

It is possible to find various apparent counterexamples to this last assertion. John Howard Griffin's book, *Black Like Me*, and Grace Halsell's *Soul Sister* are examples of individuals' reporting what happened to them when they changed their outward appearances to experience, for a while, some aspects of life as a member of a different race. There are other studies where nearly identical résumés are sent to businesses. The only difference between the résumés is an indication of the race of the person applying for the job. Both of these are instances where some aspect of race was manipulated for a real or hypothetical individual. These are experimental treatments, there

is no doubt about that. Their relevance to the use of race in social science research is, however, almost nil. Self-reported racial categories used to define a variable in a regression analysis are very different from these purported counterexamples.

On the contrary, these examples show how complex the manipulation of race really is. Grace Halsell may have changed the color of her skin, but by her own admission she could not change the fact that she was raised a southern White woman, with all of the experiences and beliefs that such an upbringing implies. In the résumé studies, it is only the race on the résumé that is changed, the altering of provided information, not the life experiences that accompany a résumé in real life. Although not entirely irrelevant, this is a far cry from changing the race of a "real" individual.

In my opinion, race does play an important descriptive role in identifying important societal differences such as those in wealth, education, and health care. The attribution of cause to race as the producer of these differences is, to me, the most casual of causal talk and does not lead to useful action.

So, relieved of the burden of raising the research status of race to that of a causal variable, I can now address the more important issue of what role race should play in causal analyses. I will discuss two related issues. The first concerns how to think about causation in racial (and other types of) discrimination. The second is how race and a true causal variable can connect in a causal study. I will illustrate this second point in more detail below in the section on biased tests.

Causation and Discrimination

If race is not a causal variable, how do we analyze issues of racial discrimination in causal terms, if at all? We certainly do think of racial discrimination in causal terms because many of us think racial discrimination is something that could be changed, reduced, or in some way altered. There are those who dream of a day when racial discrimination is a thing of the past and long forgotten. What is it that has to change? Certainly not the color of people's skin or some other physical characteristic. Clearly discrimination is a social phenomenon, one that is learned; it is taught and fostered by a social system in which it plays a complex part. When we envision a world without racial discrimination we thus envision it as a whole social system that must be different in a variety of ways from what we now see before us. One almost has to envision a parallel world, so to speak, in which things are so different that what we recognize in our own world as racial discrimination does not exist in this other parallel world. How might we detect this state of affairs in the parallel world?

I ask the reader's indulgence in my pursuing a little fantasy involving more perfect worlds that are "parallel" to our own. Something like the following might suffice to show that racial discrimination does not exist in that parallel world. Suppose we take several persons who, in the real world, have experienced what they regard as racial discrimination, and transport them into this other parallel world. There they meet their "parallel selves" and the two "selves" can exchange views about various things, including their experiences about discrimination based on race. They might have very different stories to tell each other, the parallel selves finding the stories of the original selves horrible to hear and difficult to understand from their experiences. Would that be enough to establish that racial discrimination did not exist in the parallel world? Maybe, but I think the case would be strengthened if we suppose that we also found other persons in the real world who had not had the experience of racial discrimination. Perhaps they are White, privileged, and oblivious to the plight of others? Then we transport them to the parallel world, introduce them to their parallel selves, and listen in on the resulting conversations. To put this fantasy into the simplest terms, we might then discover that the parallel selves of these privileged persons also did not report any experiences with racial discrimination.

The point of my fantasy is that racial discrimination should be viewed as how society treats different people differently in a rather complicated way. It is not just that different groups of people have different experiences, which is what statisticians would call the "main effect of race." It is the statistical interaction of race with an appropriate change in society that turns the original "different experiences" into discrimination. If discrimination were removed from society, different groups of people should experience this change differently. If instead they all experienced the change in the same way, it is hard to say, at least in my opinion, that there was ever "discrimination" in the first place.

Imagine the further complication to my fantasy if the privileged persons' parallel selves told of horrible acts of discrimination based on race. Could discrimination be said to be absent in the parallel world, or did it just get changed to some other kind of discrimination?

As one who is White and who would be considered privileged by some, I am acutely aware of how hollow-sounding a theoretical analysis of the type I have just given may appear to those on the front lines of social action. There is not much I can do about that, of course. I can only add that my intended audience are those analysts who use statistical models to estimate race effects and from them try to deduce the effects of racial discrimination. My purpose is to dissuade these analysts from using such casual interpretations of their analyses.

Race and Causes Together?

The point about discrimination being a "statistical interaction" between a (potential) change in society and racial categories of people is just a special case of a role that I think is very important for the use of race variables in analyses. Racial categories are hardly homogeneous, and treating them as such is what defines stereotyping. Yet, racial categories do capture some important phenomena that pervade many societies throughout the world. For this reason, in my opinion, the study of statistical interactions of causal variables with racial categories is a useful activity. Consider, for example, educational studies. Reading programs that are more effective for some groups of students than others are not as useful, in a general sense, as those programs whose effects are powerful throughout society. The same can be said in other domains such as medical treatments.

Whether or not racial categories are useful for finding programs that are not properly targeted for large groups of students is an empirical question. As long as wide differences in educational achievement exist between different racial/ethnic groups I am sure that checking for the interactions of program effects with race variables is both productive and easy. As I have told many a graduate student when I taught in the Graduate School of Education at Berkeley, "Please check the interactions with both gender and race of your favorite educational programs. These are two easily obtained variables and, if you find interactions with race or gender, that will tell you very interesting things about your educational program, no matter how well thought-out and implemented you think it is."

BIASED TESTS, RACE, AND CAUSE

In this final section I want to briefly integrate some of the ideas that have been put forward here in an example that combines racial categories and causes—the study of biased tests.

Claims that tests are racially (and otherwise) biased are made every day. As far as I can tell these are mostly based on the main effect of race when examining test scores. That is, racial/ethnic groups differ in their average test scores, sometimes by very large amounts—as much as one standard deviation. This "main effect of race" is not limited to one or two tests or to tests of particular formats such as multiple choice or essay. They are to be found in many tests, some would say in virtually every test.

Having been heavily involved in the study of item and test bias (Holland and Wainer 1993) I have long ago rejected the view that a simple difference in mean scores on tests or items for different groups of examinees implies that the test or items are biased. The differences in test scores between racial and ethnic groups replicate across so many tests and types of tests that ei-

ther all tests are biased or this definition makes no sense. I accept the latter rather than the former view. This is based on seeing, first hand, the extreme care that goes into the development of tests for serious uses. Indeed, the century-old application of scientific principles to test development has weeded out many sources of test bias and has made the constructs that the tests are intended to measure and the uses and consequences of the tests the paramount factors in the design and construction of modern tests.

In 1986–1987 several of us at ETS (reported in Hackett et al. 1987) developed four specially constructed "experimental" sections of a real test used for admission to a particular graduate-level course of study. We did this in order to study the effects of using item statistics to manipulate the difference in average scores between Black and White test takers. Our immediate interest was in a procedure associated with the "Golden Rule law suit settlement" (McAllister 1993). We wanted to see what effect this procedure would have on the reliability and validity of the resulting tests.

The Golden Rule procedure attempted to minimize the score differences between Black and White test takers by choosing only those test questions that minimized the performance difference between these two groups. In our study we did this but we also developed, in addition, sections of the test that maximized these differences in performance. Furthermore, we had, for comparison, other examples of the same section types for the test that had been developed in the usual professional way for this real graduate-level testing program.

The view of the proponents of the Golden Rule procedure was that by reducing the difference in the average scores of Black and White test takers, test bias was being reduced. My opinion on this is based on the observation that the performance by examinees on individual test questions varies due to many factors. In my opinion, all that the Golden Rule procedure did was to choose that subset of test items on which Black test takers performed on average somewhat higher than usual and, simultaneously, White test takers performed on average somewhat lower than usual. From my perspective both of the specially constructed types of test were biased in a sense that is clear, consequential, and as it turned out, entirely undetectable by those who only look at the words in the test booklet to assess the bias of a test.

My position is that arguments about test bias are just so many empty words unless one has examples of real tests that are biased in clearly specified ways. It is hard to say much that is useful about biased tests unless we have real examples of them for study and analysis. So the point of view that I will take here is that the two types of experimental sections were biased in favor of different groups of examinees. Some were biased in favor of Black test takers and some were biased in favor of White test takers.

Of course, following ETS test fairness rules, our experimental sections were never used in actual operational tests that affect examinee scores. They

were tested on real examinee populations but in such a way that they did not affect their reported scores. This study allow us to see if biased tests can be built to real test specifications, and, if so, how tests that are really biased behave. This work is reported in detail in Hackett et al. (1987), so I will only use a few aspects of that report to show how, in this instance, race and a causal variable worked together to give information that otherwise could not be obtained.

We used good test questions to construct our test sections. They had passed many different kinds of reviews (including those for the purpose of identifying possibly biased or "insensitive" questions) by different people and had met the usual criteria of standard statistical analyses. These were not newly developed test questions, but those that had been evaluated along the lines that serious testing programs use to produce serious tests. They were all multiple-choice questions, they all had very defensible right answers, and there was no evidence that they elicited unusual testing behavior from examinees. In my opinion, no teacher-made test in any school or university in any subject has ever been scrutinized as well as our test questions had been.

We selected two question types, Sentence Completion from the verbal dimension and Problem Solving from the quantitative dimension. These were both question types that had been used for years in the testing program in which we did our experiment. We did not introduce anything novel into the actual questions used in our study. Instead, we exploited the natural variation that occurs in actual test questions in terms of the performance on them by real examinees. Based on their pretest statistics, we grouped these questions into those that favored White examinees more than average, and those that favored Black examinees more than average. It must be clearly stated that we simply used the proportion of examinees getting each question correct as our measure of whether an item "favored" White or Black examinees. Furthermore, because of the large White-Black difference in overall performance on this nationally administered test, White examinees always averaged higher than Black examinees on each test question (i.e., the "main effect of race" mentioned earlier). Our choice of labeling of an item as biased against White or Black examinees was really a matter of how much higher the White examinees scored on it than did the Black examinees. Those questions with the smallest White/Black differences were interpreted as questions that "favored" Black examinees and those with the largest differences were interpreted as questions that "favored" White examinees. Our purpose in choosing test questions in this way was to manipulate the average score differences between White and Black examinees on the experimental test sections. This purpose was to insure that the resulting tests really did have a consequence for differences in the scores of Black and White examinees.

Our first requirement of the experienced test developers who constructed our biased sections was that they build them to meet both the content and statistical specifications that are required of any such sections for the real test. This came first because we wanted real tests not pseudo tests. Next came the biasing through the final choice of test questions using the pretest statistics as described above. As a final check, once the tests were printed we had several independent reviewers go over the sections that we had created to see if they could detect which ones were which, and they could not.

Suffice it to say we achieved all our goals. All of the test sections we had specially constructed met the content and statistical specifications for those sections. The test sections that were designed to maximize the White-Black difference in mean performance (the White-biased sections) did exactly that and the sections designed to minimize this difference (the Black-biased sections) were successful as well. Thus, we were able to create tests that varied the White-Black difference in predictable ways. In this sense, we created biased tests that were both (a) indistinguishable from the usual sections that are routinely constructed for this test and (b) that were biased in ways that could have had an impact on real scores had they been used to report real scores. They were not used, of course, in this way.

Return now to the discussion of the first parts of this chapter. What was the causal variable in this study? What we did was to arrange it so that randomly selected examinees in an operational test administration were exposed to either the White-biased sections or the Black-biased sections in a part of the operational test that did not count for their score. In addition to our special test sections, examinees also could have been randomized to one of three comparable Sentence Completion (SC) sections and to one of six Problem Solving (PS) sections. These had been constructed to meet the very same test specifications that our special test sections had been designed to meet (but not the bias, of course). These comparable sections are our control sections because they are just ordinary sections of the test developed to meet the specifications of those test sections, PS or SC. In the analysis given here, I present only the average scores over all the several control sections because they are very similar relative to the other differences that interest us.

Thus, the causal variable is the "bias type" of the section that an examinee responded to. Race will also play a role because in studying test bias we are interested in the interaction of "bias type" and race.

Table 5.1 summarizes the results of the study, emphasizing the basic messages rather than the many other relevant details that are given in Hackett et al. (1987).

The values in table 5.1 are average "formula scores," the usual raw score computed for these sections. The SC and PS sections are quite different in terms of numbers of questions and difficulty so that it is not useful to

Table 5.1. Average Section Scores for Black and White Test Takers by Subject and Type of Bias (White, Black, or Control Sections)

Subject	*Section Type*	*Black Test Takers*	*White Test Takers*
Problem Solving	Black-biased	7.2	10.4
	White-biased	4.3	10.6
	Control Average	5.5	10.1
Sentence Completion	Black-biased	10.7	12.4
	White-biased	8.6	14.3
	Control Average	9.8	13.1

compare the values across the two subjects. In case the reader is concerned that these score changes are not large enough to make a difference, I report that the standard deviation of the control sections for PS was 4.3 and for SC it was 4.4. Thus the differences between the mean scores on the Black- and White-biased sections for a given group was as large as two-thirds of a standard deviation—that is, for Black test takers on the PS sections.

I think there are three messages of table 5.1. The first is the obvious one that for this test, like many others, there is a noticeable main effect of race; that is, regardless of the type of bias used to construct the tests, White test takers score higher on average than Black test takers. Secondly, we were able to impact the scores of Black and White examinees in predictable ways using these specially constructed test sections. White scores go up (relative to the controls) for the White-biased sections and Black scores go up (relative to the controls) for the Black-biased sections.

Thirdly, and what is even more interesting to me, is that the two subjects (SC or PS) seem to behave in different ways in how the bias works. For the Problem Solving sections, the scores of White examinees are not influenced very much by the manipulation of bias type, but those of Black examinees are. However, for the Sentence Completion sections we seem to have a case of robbing Peter to pay Paul. In this case, when White scores go up, Black scores go down; and when Black scores go up, White scores go down. It could well be argued that in the PS sections the manipulation did, in fact, reduce bias for the Black test takers. But this is harder to argue for the SC sections, where some sort of "exchange" took place. In my opinion, this difference in the effect of biased tests on the scores of examinees of different races is an important point to understand with further research. Is it specific to different content areas, or question types or are there other factors involved? These are questions that can be studied and they can inform notions of test bias in ways that go well beyond the usual speculations of question wording, and such.

Once one has examples of tests that are really biased for and against different groups (rather than examples of tests that are called biased due to

their main effect of race) we can begin their scientific study with the ultimate aim of understanding how to make real tests as fair as they can be.

CONCLUSION

Race is not a causal variable and attributing cause to race is merely confusing and unhelpful in an area where scientific study is already difficult. The useful causal role of race is its ability to reveal varying effects of interventions on different parts of a diverse population. In the study of test bias, it is crucial to study the interaction of race with tests developed to have different types of bias, rather than to call the main effect of race on a single test evidence of its bias for or against different groups of examinees.