RESEARCH
REPORT

January 2003
RR-03-03

Causation and Race

Paul W. Holland

Research &
Development Division
Princeton, NJ 08541

Causation and Race

Paul W. Holland

Educational Testing Service, Princeton, NJ

January 2003

Abstract

Race is often viewed as a causal variable and “RACE effects” found from regression analyses are sometimes given causal interpretations. I argue that this is a mistaken way to proceed. RACE is not a causal variable in a very important sense of the word, and yet it does have a significant role in causal studies. The key role that RACE can play is to help our understanding of the effects of causes or interventions through the statistical “interaction” of RACE with causal variables, rather than as the main effect of RACE. These ideas are briefly illustrated using data from a study of tests constructed to manipulate the distribution of scores of Black and White test takers.

Acknowledgements

This paper is based on a presentation made at the Seminar on Racial Statistics and Public Policy at the University of Pennsylvania on January 25, 2002. The opinions expressed here are those of the author and do not necessarily represent those of his employer, Educational Testing Service, or the University of Pennsylvania. I wish to thank Carol Dwyer, Howard Taylor, and Kim Fryer for their helpful comments on an earlier draft of this report.

Introduction

For more than 2,000 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 & Palmer, 1959) and Sir A. Bradford Hill (1965); and in social science research and program evaluation there are Campbell and Stanley (1966). 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 first to consider the other *noncausal* explanations and *eliminate them.*

Two of the most commonly occurring alternatives to causal explanations are *reverse causation* and *common causes*. Here are two examples that are easy to explain and 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 employing repetitive drill and practice) are associated with *lower* student performance on NAEP. The causal explanation is that these are bad practices; they 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 for, not of the result of, these practices. The noncausal explanation is *reverse causation* in the sense that the apparent effect, i.e., low test scores, is actually a measure of the cause—that students with different needs are being taught in different ways—rather than being an effect of these practices.
- *Example 2 (Common Causes).* It is equally easy to find NAEP data that show 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 (and 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. Common causes are often called "hidden causes" or "confounders."

These examples are only two from a long list of the problems with causal explanations, but they are easy to identify and understand. Both are special cases of Simpson's Paradox (Simpson, 1951; Pearl, 2000), which has perplexed users of statistics for more than 100 years. 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 acceptance rate 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 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, & 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 to the original association.

While, I have no direct proof of this, I think Simpson's Paradox could lie behind that most curious of put-downs, "Sir, there are three kinds of lies: lies, damned lies and statistics," which is attributed by Mark Twain to Queen Victoria's prime minister, Benjamin Disraeli. One way to understand the three kinds of lies is as the blustering reaction of a great politician to someone's (mis)use of statistics to trash one of his pet policies. Versions of 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 19th century. Who knows what was the "third variable," 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 uttering naught but pure bluster if it was Simpson's Paradox he was up against. British statisticians K. Pearson and G. Yule were the first to understand the workings of the paradox but did so only years *after* Disraeli had left office.

However, the topic of this paper is not about the misplaced causal thinking that reverse causation, common causes, and Simpson's Paradox exemplify. My interest here has concerned

me for some time (Holland, 1988b) and it is of a different order than the simple misidentification of association for causation. The problem I wish to address is: What is the proper causal role of variables such as RACE and GENDER in social science research?

Every day, some economist, sociologist, or political scientist runs a regression analysis in which a variable, RACE, 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, these same analysts interpret the coefficient of RACE 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 a RACE variable *can* play some type of important role in 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 paper, I will give the details of my argument and point of view and illustrate it in the last section with an example of "biased tests."

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 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 naïve 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 paper will not contribute. When I am using "race" to mean a variable in a data analysis, I will capitalize it, as I have done above, in the manner that many computer programs do, i.e., RACE.

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) and the references therein. 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).

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 20th century invention of the sample survey employing random selection. Careful observation and description, however, have ancient scientific credentials. It sometimes helps to classify questions in terms of the interrogatives in English. The questions most relevant to descriptive studies are: *who*, *what*, *where*, and *when*. I will return to the other interrogatives shortly.

The slippery slope toward causation: Description often results in *comparisons,* and a comparison almost irresistibly invites us to use other interrogatives, in particular, *why* and *how*. These are causal questions, and, in some sense, they are more fundamental than those related to description. But, as I remind myself regularly, *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 in surveys like NAEP.

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 phenomenon, 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 catalogue 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 aware of and wary of the slide from description to causation.

Again, the interrogatives of English can begin to help us (though they have limitations). The questions of *why* and *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, occasionally, 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 questions of why. 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?" is better used for 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?" "Will making schools accountable for student test performance improve student learning?" 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.

The simplest, *ideal study* that addresses a what–if-question 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. In such a simple study *causal attribution* is easy and direct. 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 another 20th century invention, 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, people, etc.—is a fact of scientific life. In my opinion, randomized experiments are very good at addressing what–if-questions, but, by themselves, they are often not as satisfactory when it comes to the causal attributions required by questions of "Why?" and "How?" I will return to this point again later.

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 and the latter causal) and should not 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 smoke rings 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 often involved in that hallmark of science, *prediction*. Describing causal mechanisms, like identifying causes, must involve an element of speculation—sometimes a very healthy dose of it.

The description of a causal mechanism (How?) can be completely wrong while at the same time the effect of the cause (What if?) is clear and replicable. A well-known medical example concerns taking aspirin to reduce one's risk of a heart attack. The data on the risk 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 properties or its inflammation reduction properties? 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 (i.e., 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 the 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 careful

assessment of effects of interventions, as well as plausible explanations of why they occur, are *both* important, but they need to be clearly distinguished and not confused with each other.

Causal variables. It should be clear, but 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, or other causal study, 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, 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 & Singer, 1988), but I still think it is a sound position and reiterate it here.

Our slogan closely corresponds to the simple, ideal study, already mentioned, for understanding causation. Two identical/similar units of study, one exposed to one experience and the other exposed to another experience, are subsequently compared on an identical/similar outcome criterion. In this basic setting, the attribution of cause refers 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 was obtained under the experience the unit was actually exposed to and (b) the result that would have been 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, we usually have no problem identifying the causal variables. (Even there, however, what was actually manipulated may not be so clear—perhaps the most famous examples being those involved with comparisons to placebos. Is it the effect of the drug or of just taking a pill?)

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 individuals 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 medical action is concerned it often does not matter whether the association is causal or noncausal. In medicine, the term “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 non-causes are often not. For example, the physician can help you to stop smoking, but not to 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 has enduring value and not its status as a causal variable.

Is RACE a Cause?

From the arguments in 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 studies where nearly identical resumes are sent to businesses. The only difference between the resumes is an indication of the RACE of the person applying for the job. In both these studies and these books, 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 RACE in a regression analysis are very different from these purported counterexamples.

These examples show, instead, 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 resume studies, it was only the RACE on the resume that was changed—altering provided information—not the life experiences that accompany a resume in real life. Although not entirely irrelevant, this is a far cry from changing the race of a "real" individual.

In my opinion, RACE can 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 variables can 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 in the last section.

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. Some dream of a day when racial discrimination is a thing of the past and long forgotten. What 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, taught, and fostered by a social system in which it plays a complex part. When we envision a world without racial discrimination, we 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 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 we transport them into this other world. There they meet their parallel selves and exchange views about various things, including their experiences with 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. 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 to 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 *difference* in society that makes the original different experiences into *discrimination*. It is not just that things are better for everyone in the parallel world (that is a "main effect"), but that there is a difference for some parallel selves and not for others. If discrimination were removed from society, different groups of people should experience this change differently. If, instead, they all experienced the difference in the same way, I would find it hard to say that there was ever "discrimination" in the first place. We don't call it discrimination that children can't vote, but adults can. It is discrimination when only some children can vote when they become adults.

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

As one who is White and who would be considered privileged by some, I am acutely aware of how hollow sounding such a theoretical analysis might appear to those on the front lines of social action. I can't do much about that, of course. I can only add that my intended audience is those analysts who use statistical models to estimate RACE effects and from the results try to deduce the effects of racial discrimination. My purpose is to dissuade these analysts from using such casual causal interpretations of their analyses.

RACE and Causes Together?

The point about discrimination being a "statistical interaction" between a (potential) difference in societies 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. In this section I will discuss this role more carefully.

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 RACE variables 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 felt 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 I have put forward here in an example that combines a RACE variable 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 on 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, and some would say in virtually every test.

Having been heavily involved in the study of item and test bias (Holland & Wainer, 1993), I 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 either all tests are biased or this definition makes no sense. I accept the latter rather than the

former view. This is based on seeing, firsthand, 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.

From 1986 to 1987, several of us at ETS (reported in Hackett, Holland, Pearlman, & Thayer, 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 empirical performance difference between these two groups. In our study, we did use this procedure, but we also developed 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. I differ on this interpretation and base my opinion 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 without having 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 allowed 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, Holland, Pearlman, and Thayer (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 in 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 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. We wanted 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 in the first parts of this paper. 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 also will play a role because in studying test bias we are interested in the interaction of "bias type" and RACE.

Table 1 summarizes the results of the study, emphasizing the basic messages rather than the many other relevant details that are given in Hackett, Holland, Pearlman, and Thayer (1987).

Table 1

Average Section Scores for Black and White Test Takers by Subject and Type of Bias

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

Note. Type of bias = White, Black, or control sections.

The values in Table 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 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 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—i.e., for Black test takers on the PS sections.

I think there are three messages in Table 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. Second, 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.

Third, 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 PS 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 SC 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. One could argue 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 important, and to understand it, we need further research. Is it specific to different content areas or question types or are 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, etc.

Once we have 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 or other variables), we can really begin the scientific study of test bias. Such research can strengthen the many other efforts of professional test developers towards making *real* tests as fair as they can be.

I need to make a final comment about my use of the term "bias." In reviewing this paper, the Princeton sociologist Howard Taylor reminded me that bias is usually regarded as a difference in the way a test score predicts some important outcome variable—such as success in college as measured by GPA. If a test predicts differently for one group versus another, it is biased as a predictor. Famous examples of differential prediction abound as reported in Willingham and Cole (1997) and Willingham, Pollack, and Lewis (2000) and the references therein. This looks like a counterexample to my insistence on "bias" requiring a comparison to another test. I view it differently, of course. That a test predicts grades differently for one group than another, for example, is merely a more complicated version of what I have called here the

"main effect of RACE." Instead of just an average test score being different for different groups, it is the conditional mean grade given a test score that is different for different groups. This move from the marginal distributions of a test score to the conditional distribution of a criterion score given the test score is just added complexity but nothing more. Prediction equations can be different for different groups of examinees for a whole host of reasons. If changes in a test *change* these differences in prediction, we may have some basis for talking about bias in one test relative to another. I couched my discussion in terms of the simpler situation of changing the relative positions of mean test scores for two groups because it is easier to understand and allowed a more direct discussion of one test being biased for or against a group relative to another test.

Discrimination and bias are contentious and difficult issues. Obscuring them with simplistic calculations that do not attend to the proper role of RACE in a causal study helps no one.

Summary

RACE variables are not causal variables 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.

References

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science, 187,* 398-404.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Coleman, J. S. (1966). *Equality of educational opportunity*. Washington, DC: US Government Printing Office.

Hackett, R. K., Holland, P. W., Pearlman, M., & Thayer, D. T. (1987). *Test construction manipulating score differences between Black and White examinees: properties of the resulting tests* (ETS RR-87-30). Princeton, NJ: Educational Testing Service.

Hill, A. B. (1965). The environment and disease: Association or causation. *Proceeding of the Royal Society of Medicine, 58,* 295-300.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945-970.

Holland, P. W. (1988a). Causal inference, path analysis and recursive structural equations models. In Clifford C. Clogg (Ed.), *Sociological methodology 1988* (pp. 449-484). Washington, DC: The American Sociological Association.

Holland, P. W. (1988b). Causal mechanism or causal effect: Which is best for Statistical Science? *Statistical Science, 3,* 149-195.

Holland, P. W. (2001). The causal interpretation of regression coefficients. In M. C. Galavotti, P. Suppes, & D. Costantini (Eds.), *Stochastic causality* (pp. 173-188). Stanford CA: CSLI Publications.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Earlbaum Associates.

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

McAllister, P. H. (1993). Testing, DIF, and public policy. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 389-396). Hillsdale, NJ: Earlbaum Associates.

Marini, M. M., & Singer, B. (1988). Causality in the social sciences. In Clifford C. Clogg (Ed.), *Sociological methodology 1988* (pp. 347-410). Washington, DC: The American Sociological Association.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge: Cambridge University Press.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society B, 13,* 238-241.

Yerushalmy, J., & Palmer, C. E. (1959). On the methodology of investigations of etiologic factors in chronic diseases. *Journal of Chronic Diseases, 10,* 27-40.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

Willingham, W. W., Pollack, J. M., & Lewis, C. (2000). *Grades and test scores: Accounting for observed differences* (ETS RR-00-15). Princeton, NJ: Educational Testing Service.