14.11 Spring 2006: Putting Social Science to the Test. Lecture Note #1: Discrimination

David Autor and Esther Duflo

March 7, 2006

1 Introduction

It is useful to begin with three definitions (from Merriam-Webster):

1. The quality or power of finely distinguishing.

2. The act, practice or an instance of discriminating categorically rather than individually.

3. Prejudiced or prejudicial outlook, action or treatment.

The first of these seems to be clearly a value-neutral activity; finely distinguishing is not an objectionable activity.

The second type of behavior ('distinguishing categorically') is much more morally nuanced. For example, when I buy a new car, I might distinguish categorically between American and Japanese cars. I might be unwilling to buy an American car because I expect it will be unreliable or, similarly, I might be willing to pay extra for a Toyota. This is categorical thinking; independent of my knowledge of a specific car (the model, the individual instance), I use prior information about American and Japanese cars as a group to make purchasing decisions. Is this objectionable?

What if I use similar reasoning in some other domain? For example, it's common parenting practice to coach children (esp. girls) that if they are lost, they should ask a woman—preferably one with children, and definitely *not* a man—for help.

What if I am a taxi-driver applying the same type of reasoning when deciding whether or not to stop for a passenger? I might believe that some groups have higher probability of robbing me than others. Or, I might expect that passengers from some groups will ask me to drive to a neighborhood that I view as unsafe.

All of the above are examples of categorical thinking—what economists call 'statistical discrimination'— where I use characteristics of a group (Japanese cars, mothers with children, minority passengers) to make educated guesses about the characteristics of individual instances from these groups (a specific car, a specific mother with children, a specific minority passenger). You might also call this behavior 'prejudicial outlook' (the third definition) since my expectations about cars, genders, race groups are partly based on 'prejudgments' from prior experiences with these categories.

Using categorical information to resolve uncertainties about the properties of specific examples often seems a good idea. I cannot know the quality of every specific car for sale at the dealership (even within a specific model, there is quality variation). So, I may be wise to use information about the category to make educated guesses about the individual instance. *Does it matter if my expectations (prejudgments) are accurate on average?* (Though they will always be wrong in specific cases—not every American car is unreliable or every Japanese car dependable.)

There is another type of discrimination that is hinted at by the third definition though not made precise. This is what we would call 'animus-based' or 'taste-based' discrimination. Here, even knowing that two instances of a specific category are identical for 'utilitarian' purposes, I might have a strong preference for one or the other based on some arguably extraneous characteristic. For example, quality constant, I might be inclined *not* to buy American cars because I do not like 'what American cars stand for.' Or, I might not stop for minority passengers in my taxi cab because I hold a racial or ethnic animus.

This type of discrimination is fundamentally different from the 'statistical' form because it is based only on taste, *not* on uncertainty. Of course, not all tastes are objectionable. You can discriminate against chocolate ice cream in every U.S. state and no one will be offended. But certain tastes are viewed as illegitimate. It is generally considered illegitimate to harbor (or, at least to act upon) a distaste for minorities or women (though this was clearly not always true).

It is therefore useful to distinguish between two conceptually different forms of discrimination. Animus-based ('taste-based') discrimination and statistical ('educated guess' or 'information-based') discrimination. Both may be problematic, but they stem from different motivations, have different empirical implications, and potentially could be resolved using different policies. (Arguably, statistical discrimination is less objectionable but also harder to eliminate.)

We'll discuss taste-based and information-based models of discrimination and consider some experimental evidence on their importance. We'll also consider a third type of reaction to discrimination that does not comfortably fit under either definition. This is so-called 'stereotype threat.'

Before beginning, we should note that the economics literature on discrimination has its roots in *employment* discrimination. This in part reflects who wrote the early literature (it's originator, Gary Becker, is a labor economist) and partly because a key focus of public debate on discrimination in the 1960s and 1970s was about employment. Much (not all) of our discussion below will also be couched in terms of employment.

2 Models of discrimination

2.1 Taste-based discrimination

An enormous literature, starting with Becker's 1957 book *The Economics of Discrimination*, explores (surprise!) the economics of discrimination. Economic models of discrimination can be divided into two classes: competitive and collective models. Competitive models study individual maximizing behavior that may include discrimination. In collective models, groups act collectively against each other. Almost all economic analysis has focused on competitive models, and we'll do likewise here. Becker's model studied taste-based discrimination. We'll start with that.

In formal economic terms, discrimination is when members of a minority group are treated differently (less favorably) than members of a majority group with identical productive characteristics. Let the wage Y be the wage equal to

$$Y_i = X_i\beta + \alpha Z_i + e_i, \tag{1}$$

where X_i is a vector of exogenous productivity characteristics and Z_i is an indicator variable for membership in a minority group. Assuming that $X_i\beta$ fully captures the set of productive characteristics and their returns and/or Z_i is uncorrelated with e, then discrimination is a case where $\alpha < 0$.

We already face three difficulties just using this simple definition.

1. 'Productivity' may directly depend on Z—for example, in the entertainment industry or a market in which customers value Z in workers (e.g., discriminating customers). If customers will pay more to see a white actress or a black athlete, is this a legitimate component of productivity?

2. There is also a question of whether β—the production technology—is truly exogenous. For example, operating fire fighting equipment requires considerable physical strength and stature. This has historically been used as an argument against the entry of women into this profession. But these physical requirements are engineered attributes and probably could be altered. If humans were 20 percent less physically strong, presumably they could still fight fires. It seems likely that fire-fighting equipment used in Japan has historically demanded a smaller physical stature.

3. The $X's$ could also be endogenous. Pre-market discrimination—or expectations of future discrimination—could reduce $X's$ for members of the minority group. (Examples: poor schools, or a rational belief among minorities that education will not be rewarded by the market.)

Point (1) is one we may be able to examine directly. Point (2) and (3) are much harder to test. But whether or not these are relevant, it can still be the case that $\alpha < 0$ conditional on both X and β, which would constitute discrimination.

In Becker's 1957 model, employers have a 'taste for discrimination,' meaning that there is a disamenity value to employing minority workers. (Hence, discrimination comes directly out of the utility function.) In this case, minority workers may have to 'compensate' employers by being more productive at a given wage or, equivalently, by accepting a lower wage for identical productivity. The basic insights of this model require almost no formalization. We will formalize only very slightly.

- Let A denote majority group membership and B denote minority group membership.
- Employers will maximize a utility function that is the sum of profits plus the monetary value of utility from employing members of particular groups.
- Let d be the taste parameter of the firm, which Becker called the "coefficient of discrimination."
- Firms will maximize

$$U = pF\left(N_b + N_a\right) - w_a N_a - w_b N_b - dN_b,$$

where p is the price level, F is the production function, N_x is the number of workers of group $x = \{a, b\}$, and w_x is the wage paid to members of each group.

- Employers who are prejudiced ($d > 0$) will act as if the wage of b group members is $w_b + d$. Hence, they will only hire b group members if

$$w_a - w_b \geq d.$$

- Let $G(d)$ denote the cumulative density function (CDF) of the prejudice parameter d in the population of employers.

- The optimal number of workers hired at each firm is determined by the solutions to

$$\begin{aligned} pF'(N_a) &= w_a, \\ pF'(N_b) &= w_b + d. \end{aligned}$$

- Treating p as fixed and aggregating across firms in the economy leads to the market demand functions $N_a^d(w_a, w_b, G(d)), N_b^d(w_a, w_b, G(d))$ for each worker type. Wages are determined by

$$\begin{aligned} N_a^d(w_a, w_b, G(d)) &= N_a^s(w_a), \\ N_a^d(w_a, w_b, G(d)) &= N_a^s(w_b), \end{aligned}$$

  where $N^s(\cdot)$ are the supply functions for the worker types.

- Notice the main point that comes out of this setup is this: A wage differential $w_b < w_a$ will arise if and only if the fraction of discriminating employers (or discriminating jobs) is sufficiently large that the demand for B workers when $w_b = w_a$ is less than the supply.

- In other words, *discrimination on average does not mean discrimination at the margin.* If there are enough non-discriminating employers, then discrimination is competed away. This also implies that minority workers don't work for discriminating employers.

- If, however, the share of prejudiced employers is sufficiently large, then some b group members will work at $d > 0$ employers, and this implies that $w_b < w_a$. In this case, the

strength of prejudice at the margin (that is d for the marginal employer of b workers) is what determines the size of the wage gap.

- With free entry or constant returns to scale (CRS), discriminating employers may be competed out of business. In a competitive market, each worker must earn his marginal product. Under CRS, non-discriminating firms would simply expand to arbitrage the wage differential borne by minority workers. In equilibrium, discriminating employers must fund the cost of their distaste out of their own pockets; they cannot pass the cost onto the minority worker.

So, to summarize:

- In partial equilibrium, minority workers must 'compensate' employers by being more productive at a given wage or, equivalently, accepting a lower wage for equivalent productivity.

- These tastes create incentives for segregation. It is potentially Pareto improving for minority workers to work in their own businesses and similarly for majority workers—then no one bears the cost of the distaste.

- In general equilibrium, these tastes can only be indulged at a positive cost to the employer.

Key testable implications of this model are:

- 1. Wage differentials: Minority workers earn less than majority workers of identical productivity.

  2. Preferential hiring: Employers are less likely to hire minority workers of identical productivity.

- But these implications may not apply in equilibrium—so it's not clear when (if ever) we should observe them.

2.2 Statistical Discrimination

Much economic analysis of discrimination since Phelps (1972) and Arrow (1973) has focused on the statistical theory of discrimination rather than on taste-based discrimination. The premise of the statistical discrimination literature is that firms have limited information about the skills of job applicants but hold no animus against racial groups.

Statistical discrimination is the solution to a signal extraction problem. If an employer observes a noisy signal of applicant productivity and also has prior information about correlates of productivity (let's say a group-specific mean), then the expectation of applicant productivity should place weight on both the signal and the mean (in fact, both are 'signals.').

The mathematics of this model are slightly more involved than the one above and we won't do the formal development in 14.11 (if interested, ask your instructors for more details). The intuition of the model is straightforward, however. It says that a person (such as an employer) should normally place *some* weight on prior information about a category when attempting to draw an inference about a specific instance of that category so long as the information about the specific instance is not very precise (i.e., so I should use my knowledge of Toyotas generally to evaluate whether the new 2007 Toyota Prius is likely to be reliable).

2.2.1 Statistical discrimination: Efficiency, Legality, Fairness,

Efficiency

1. Unlike taste-based discrimination, statistical discrimination is not competed away in equilibrium. So, we can be reasonably confident that, if it exists, we should be able to detect it in a general set of cases.

2. Closely related to (2), statistical discrimination is 'efficient.' That is, because statistical discrimination is the optimal solution to an information extraction problem, economists might generally say that employers *should* statistically discriminate. It is profit-maximizing, it is not motivated by animus and it is arguably 'fair' since it treats people with the same *expected* productivity identically (though not necessarily with the same *actual* productivity). Hence, many economists might endorse statistical discrimination as

a reasonable public policy.

Legality It is illegal in the United States to make hiring, pay or promotion decisions based on predicted performance if predictions are based on race, sex, age, disability or union membership. (An employer presumably can statistically discriminate among non-disabled, white males under age 40.) Statistical discrimination is probably difficult to detect, so it is plausible that it occurs frequently (perhaps even unintentionally) despite the law.

Fairness Legality aside, it is worth asking whether statistical discrimination accords with intuitive notions of fairness. Let's take a loaded example: racial profiling. Say you are the New Jersey State Police and that drug runners travel on your highways. You have limited resources to expend on stopping cars and (of course) you want to use these resources efficiently. You know as a proven statistical pattern that cars driven by Latino drivers are more likely than average to be running drugs. All else equal, should you be more inclined to pull-over cars driven by Latino drivers?

2.3 Stereotype Threat

The 'stereotype threat' hypothesis originates with psychologist Claude Steele and coauthors. This hypothesis says that members of groups that are 'stereotypically' believed to have negative attributes may behave in ways that confirm these attributes when the 'stereotype threat' is made salient. For example, female mathematicians may perform badly on math tests when they are reminded that many people believe that women are not as capable as males at mathematics. Or, blacks may perform poorly on IQ tests when the tester subtly suggests that blacks are not as intellectually capable as whites.

This hypothesis may sound far-fetched, but in fact it is easy to think of cases where it *could* be relevant. Steele gives the example of a black male sociologist who feels anxious when he is waiting in line at an ATM if the customer ahead of him happens to be a white woman. Although this sociologist has no criminal intent, he is aware that white female ATM customers may be made anxious by his presence, believing that black males pose a criminal threat. Presumably, the sociologist reacts to this 'stereotype threat' by trying to appear especially non-threatening,

perhaps by keeping exaggerated physical distance from other customers. (It's conceivable that, opposite of the intention, this makes other customers more nervous.)

In this example, the stereotype threat does not make the sociologist more likely to engage in a criminal act (which the Steele hypothesis might suggest it should). But it does support the idea that members of discriminated groups may be acutely aware of stereotypes—at least in situations that make these stereotypes salient—and that this awareness could potentially affect behavior and outcomes. (For a fictional example in which stereotype threat experienced by black males *does* directly lead to a criminal act, see the car-jacking scene in the 2005 movie *Crash*.)

Anecdotes are not evidence, but, as we will see, the experimental evidence favoring the stereotype threat hypothesis is somewhat remarkable.

Note that the stereotype threat hypothesis does not fall under the other two categories. It is neither animus-based nor statistical discrimination; it is self-fulfilling prophesy. One could write an economic model about this, but it would be difficult in such a model to motivate the idea that an individual would choose to behave in a way that is self-destructive in response to a stereotype.

3 Causal inference in social science

You are advised to read the 1986 JASA article by Holland to get a deeper understanding of this material.

Before we begin discussing experimental results, we need to consider why we bother to experiment at all. Experiments are a lot of work. Why not simply measure correlations and save ourselves the effort of experimenting? Much of social science (psychology, anthropology, sociology, political science, epidemiology and large parts of economics) concerns analyzing correlations among variables—i.e., the correlation between education and income, the correlation between obesity and heart disease, the correlation between happiness and longevity.

Correlation describes the statistical relationship between two observed variables. Correlation has no necessary relation to cause and effect. You can measure the correlation between happiness and longevity with great precision and yet know nothing about how making someone

happier affects their longevity (maybe the same gene that causes longevity causes happiness, so making someone happier would not increase their longevity.)

In this class, we are not generally interested in these correlations. Our goal is to analyze causal questions. Why? Because science advances through analyzing cause and effect relationships, not (primarily) by documenting correlations.

Causal questions:

- What is the effect of race on the odds of getting a job offer?
- What is the effect of happiness on longevity?
- What is the effect of growing up in a poor neighborhood on criminal behavior?

These questions are much harder to answer than the correlational questions. Correlations are readily measured from observational data. Causal effects can *never* directly be measured (to be explained). Two things to bear in mind:

1. **A causal question intrinsically concerns comparing a 'factual' (actual) state to a counterfactual state that (by definition) has not occurred.** When we say, what is the causal effect of Z on Y, we mean, what value would Y have taken if Z were some other value? It's easiest to phrase this for a binary cause, $Z \in \{0, 1\}$. So, if we know that $Z = 1$ and $Y = Y_1$, the causal question is, what would Y have been if $Z = 0$ instead. That's a question that involves *rolling back time* to experience a different counterfactual reality.

2. **A causal effect is always intrinsically measured relative to some alternative cause.** It is not meaningful to ask what is the causal effect of Z on Y without (at some level, perhaps implicitly) specifying what alternative Z we are comparing it to; the counterfactual Y depends on the counterfactual Z. For a binary case, we want to compare Y_1 to Y_0.

Some notation:

Let Y_i be the outcome of interest for unit i, where i could be a person, a cell, a drop of water, a sovereign country. Let's suppress the subscript i where possible.

We want to consider two possible outcomes for i. Let Y_0 be the outcome if $Z = 0$ and Y_1 be the outcome for $Z = 1$.

Thus, for every unit i, we can imagine two potential outcomes $\{Y_0, Y_1\}$ that we would observe if the unit were treated ($Z = 1$) or untreated ($Z = 0$). Although we only observe either Y_0 or Y_1 (never both), we assume that both are well-defined.

The causal effect of Z on Y is therefore $T = Y_1 - Y_0$ (where T stands for Treatment Effect).

The problem that this immediately reveals is that *we never observe* $Y_1 - Y_0$ for a single unit i.

Instead, we observe $Y = Y_1 Z + Y_0 (1 - Z)$.

Fundamental Problem of Causal Inference: It is not possible to observe the value Y_1 and Y_0 for the same unit i and so we cannot measure the causal effect of Z on Y for unit i.

You may say: Why can't we just switch Z from 0 to 1 and back to observe both Y_1 and Y_0? In fact, this procedure is not informative about $Y_1 - Y_0$ without further assumptions (discussed below).

Solutions to the Fundamental Problem of Causal Inference?

1. We may assume *temporal stability* and *causal transience* ('Laboratory assumptions'). If the causal effect of Z on Y is the same at any point in time (now, the future) *and* the causal effect of Z on Y is reversible (so having once exposed Y to Z doesn't permanently change the effect of Z on Y), then we can observe $Y_{1i} - Y_{0i}$ simply by repeatedly changing Z from 0 to 1. Formally, these assumptions are: $Y_{1it} = Y_{1i}$, $Y_{0it} = Y_{0i}$ for all i and t where t indexes time. Notice that temporal stability and causal transience are *postulates*; they cannot be directly tested.

   Example: You can turn water from ice to steam and back repeatedly to analyze the causal effect of temperature change on water molecules. But what allows you to draw a causal inference that steam is the counterfactual for ice when the treatment is 100 versus 0 degrees Centigrade is the belief that water molecules are not fundamentally altered by

heating and cooling, and that the relationship between temperature and the behavior of water is stable (e.g., does not depend on the phase of the moon).

But would it be valid to assess the effectiveness of a cancer treatment on patient i by repeatedly administering the treatment, testing for cancer, withdrawing the treatment, testing for cancer, etc.?

2. We may assume *unit homogeneity* ('Homogeneity assumption'). If the Y_{1i} and Y_{0i} are identical for all i, then we can measure the causal effect simply by calculating $Y_{1i} - Y_{0j}$ for $i \neq j$. Again, unit homogeneity isn't normally a verifiable assumption. But for laboratory conditions, it might be reasonable (e.g., experimenting with two molecules of water). This would clearly be invalid for two cancer patients.

3. Statistical solution (different from the other two):

   - For human subjects, neither temporal stability, causal transience nor unit homogeneity will ever hold. So, let us start by acknowledging that we cannot observe $T_i = Y_{1i} - Y_{0i}$.
   - We would potentially be happy to settle for some kind of population average treatment effect instead:

     $$\hat{T} = E\left[Y_1 - Y_0 | Z = 1\right] = E\left[T | Z = 1\right],$$

     where $E\left[\cdot\right]$ is the 'expectations' operator, denoting the mean of a random variable. This expression above defines the effect of 'treatment on the treated,' that is the causal effect of the treatment on the people who receive it ($Z = 1$).
   - Since we cannot directly observe T for any individual i, how do we estimate $E\left[T | Z = 1\right]$?
   - One idea: We could compare $E\left[Y | Z = 1\right]$ and $E\left[Y | Z = 0\right]$ to form $\tilde{T} = E\left[Y | Z = 1\right] - E\left[Y | Z = 0\right]$. Is this a good idea? For example, let Z be the cancer treatment and $Y \in \{0, 1\}$ be a binary variable denoting cancer diagnosis. We could compare cancer rates among those taking the treatment ($E\left[Y | Z = 1\right]$) versus those not taking the

treatment ($E[Y|Z=0]$) to estimate the causal effect of the treatment on cancer incidence.

- The cancer example should make it intuitively clear why $\tilde{T}$ is *not* a good estimator for $E[T]$. The problem is that there is no reason to assume (and many reasons to doubt) that $E[Y_1|Z=1] = E[Y_1|Z=0]$ and $E[Y_0|Z=1] = E[Y_0|Z=0]$. Calculating $E[Y_1|Z=1] - E[Y_0|Z=0]$ will not generally give us $E[Y_1 - Y_0|Z=1]$. The underlying problem is that it is very unlikely that Y_1, Y_0 are independent of Z.

  More concretely, people who obtain a cancer treatment are generally not comparable to those who don't (in particular, the former group is likely to have cancer!). So, $E[Y|Z=1|] - E[Y|Z=0] > 0$. So, if we estimated $\tilde{T} = E[Y|Z=1] - E[Y|Z=0]$, we would almost surely conclude that treatment Z *causes* cancer, even if Z is effective. The problem is that $E[Y_1 - Y_0|Z=1] \neq E[Y_1|Z=1] - E[Y_0|Z=0]$. More precisely, it is nearly certain that $E[Y_1|Z=1] > E[Y_1|Z=0]$ and $E[Y_0|Z=1] > E[Y_0|Z=0]$, i.e., people who take the cancer treatment are more likely to have cancer than those not taking the treatment *regardless* of whether they take the treatment. This is *not* because the treatment causes cancer but because having cancer causes people to seek treatment. Hence, a simple comparison of the cancer rates of those taking and not taking treatment is uninformative about the causal effect of the cancer treatment on the cancer rate of any individual or population.

  Formally:

$$E[Y_1|Z=1] - E[Y_0|Z=0] = \underbrace{E[Y_1|Z=1] - E[Y_0|Z=1]}_{T} + \underbrace{\{E[Y_0|Z=1] - E[Y_0|Z=0]\}}_{Bias}.$$

  The first term on the right-hand side of this equation is the true, causal effect of the cancer treatment on those who take it (the effect of 'treatment on the treated'), and the second term is the potential bias that occurs if the counterfactual (non-treated) outcomes of the control group differ from the counterfactual (non-treated) outcomes of the treatment group.

Another example [for self-study]

Let's say Y is the number of mathematical expressions you can differentiate in an hour after 4 years of college and Z is an indicator variable for whether or not you attended MIT. If we administered math tests at random, we would certainly find that $\tilde{T} = E[Y_1|MIT = 1] - E[Y_0|MIT = 0] > 0$, i.e., MIT students do more calculus in an hour than non-MIT students.

Is $\tilde{T}$ a valid estimate of the causal effect of attending MIT on calculus skills (that is, a valid estimate of $E[Y_1 - Y_0|MIT = 1]$? No. Students who are skilled in calculus choose to come to MIT, and they would be more skilled than the average student in calculus, regardless of whether they attended MIT. So, $E[Y_0|MIT = 1] > E[Y_0|MIT = 0]$. That is, students who attended MIT would have done better at math than students who didn't MIT even if the attendees had *not* attended MIT. So $\tilde{T} > T$. [See expression above.] We do not obtain a valid causal estimate of the effect of attending MIT on math skills of the students who attended MIT by comparing them to students who didn't attend MIT.

The substantive problem (again) is that attendance at MIT is endogenous. Students come to MIT in part because they are good at math. It is unreasonable to assume that non-MIT students are a valid comparison group for MIT students.

- Let's say instead that we picked a large number of $i's$ at random and randomly assigned half to MIT= 1 and half to MIT= 0. This pretty much guarantees (unless we are very unlucky) that $E[Y_1|MIT = 1] = E[Y_1|MIT = 0]$ and $E[Y_0|MIT = 1] = E[Y_0|MIT = 0]$. Consequently:

  $$\hat{T} = E[Y_1 - Y_0] = \underbrace{E[Y_1|MIT = 1] - E[Y_0|MIT = 1]}_{T} + \underbrace{\{E[Y_0|MIT = 1] - E(Y_0|MIT = 0)\}}_{bias\ =0}.$$

  In this case, randomization removes the bias term by ensuring that

  $$\{E[Y_0|MIT = 1] - E[Y_0|MIT = 0]\} = 0.$$

In summary, randomization potentially solves the causal inference problem by making the treatment status $Z = \{0, 1\}$ independent of potential outcomes: $E(Y_1), E(Y_0) \perp Z$ meaning $E[Y_1|Z = 1] = E[Y_1|Z = 0]$, $E[Y_0|Z = 1] = E[Y_0|Z = 0]$. This observation motivates the idea of using a randomly selected 'control group' to ensure that the group not receiving the treatment provides a valid estimate of the *counterfactual* outcome for the treated group.

To solve the Fundamental Problem of Causal Inference in Economics, we almost always use the statistical solution. This is because it is rarely plausible for human behavior that either of the two alternative solutions applies (temporal stability + causal transience or unit homogeneity). By contrast, the statistical solution is almost certain to work.

4 Evidence on discrimination

We will consider evidence on 'types' of discrimination discussed above: animus-based, statistical and 'stereotype threat.'

4.1 A Quasi-Experiment in the Labor Market: Orchestrating Impartiality, Goldin and Rouse (AER, 2001)

This creative study attempts to isolate the importance of gender preference in a market setting: orchestra auditions. Very simple idea: During the 1970s since 1990s, some orchestras started using screens during solo auditions to hide the identity of performers. Women were historically viewed as unsuitable for orchestras. Did the use of blind screens improve their chances of getting a job?

Statistically, this study uses a 'differences-in-differences' design. The authors evaluate the success rate of women relative to males at orchestras using blind auditions relative to orchestras using non-blind auditions. Using the same group of subjects (individual performers) observed in both venues makes this comparison informative.

Let's develop this logic formally. We would like to know the effect of a candidate's gender on her probability of hire. You may be tempted to think that the causal effect of interest is:

$T_i = E\left[Y_{1i} - Y_{0i}|F_i = 1\right]$, where $F = 1$ indicates that the candidate is female ($F = 0$ indicates male), Y_{1i} is the probability of hire if female, and Y_{0i} is the probability of hire if male. But this framing of the question doesn't quite make sense. We cannot actually manipulate the gender of an applicant (without changing many, many other things), so it's perhaps not meaningful to ask how an applicant i would have fared were she instead a male.

But there is a way to frame the question that does not run afoul of this 'reality' constraint. Consider instead asking the question: how would Male and Female applicants have fared under a blind audition system (i.e., where gender is not known) relative to a non-blind system where gender is revealed. This is a reasonable alternative question to ask because we believe that the *relevant* criterion for selection should be independent of information on gender. Because we believe that the quality of performance is the only relevant productivity criterion, is it interesting to ask whether masking versus revealing gender affects hiring.

How shall we formally frame this question? Let Y_1 equal the outcome of a person in a blind audition and Y_0 equal the outcome under a nonblind audition. In this case, the treatment effect of interest is $T = E\left[Y_1 - Y_0|B = 1\right]$, that is the difference in outcomes for a women auditioning under a blind screen relative to the outcome she would have experienced under a nonblind screen.

That's a step in the right direction, but there are several issues that we need to consider here before moving to the data:

1. Blind versus nonblind auditions may have a direct effect on hiring odds for both males *and* females. Thus, it may be misleading to only estimate the blind/nonblind contrast for females.

2. The women who choose (or are selected) to audition for orchestras using blind screens may be different from those who choose orchestras with nonblind screens.

3. It's finally possible that performers perform differently in front of blind and nonblind screens—they could be more or less nervous.

The first two of these problems, we can handle. The third we cannot solve using the data available to Goldin and Rouse. We'll discuss why in a minute.

Let's write the expected probability of hire for a performer in a non-blind and blind audition as:

$$\begin{aligned} E[Y_{0i}] &= \alpha_i + \gamma F_i, \\ E[Y_{1i}] &= \alpha_i + \beta. \end{aligned}$$

Thus, the hiring probability in the non-blind condition is a function of individual ability (α_i) and possibly a gender discrimination coefficient (if $\gamma < 0$). The hiring probability in the blind condition is a function of individual ability and a blind condition 'main effect' (β) which may affect hiring odds for both genders. [Note, there is no reason to also introduce a non-blind main effect since the blind main effect is really only defined as a contrast to the non-blind condition].

So, if we contrast the outcomes of females ($F = 1$) who audition at both blind and non-blind conditions, we obtain

$$E[Y_{1i} - Y_{0i}|F_i = 1] = \alpha_i + \beta - \alpha_i - \gamma = \beta - \gamma.$$

Thus, this contrast gives us a combination of the causal effect of interest and a nuisance parameter, which is the main effect of the blind condition. How can we eliminate this nuisance parameter? Consider the analogous contrast for males:

$$E[Y_{1i} - Y_{0i}|F_i = 0] = \alpha_i + \beta - \alpha_i = \beta.$$

Combining these equations gives us the 'difference-in-difference' estimator

$$\hat{T}_{DD} = E[Y_{1i} - Y_{0i}|F_i = 1] - E[Y_{1i} - Y_{0i}|F_i = 0] = \gamma.$$

Thus, the difference-in-difference estimator solves *two* problems here: first, contrasting outcomes of females in blind/non-blind relative to males in blind/non-blind eliminates the direct effect of the blind/treatment on hiring odds (so we can isolate the pure gender effect). Second, using a sample of candidates who auditioned in *both* venues allows us to eliminate the pure quality effects (α_i) that might otherwise bias our estimates (e.g., if more capable women choose to participate in primarily at blind auditions).

Why does this estimation strategy not solve the problem if performers perform differently during blind and nonblind auditions? Because in that case, α_i is not constant (it will depend on whether or not the audition is blind).

Okay, now we are ready to look at the data.

- See Table 1 for a summary on the implementation of screens
- See Figure 3 for long-term trends in female hiring
- Table 4: On average, women do *worse* on blind rounds. But this could be due to composition of the female pool in blind rounds. It's possible that only the very best women compete when the game is lopsided (i.e., in the non-blind rounds).
- Table 5: Models limited to musicians (male and female) who auditioned both blind and non-blind suggest that women did relatively better in blind rounds (diff-females minus diff-males).
- Table 6 gives the main estimates.
- Table 7 estimates models for the 3 orchestras that switched policies; here the authors can include musician and orchestra fixed effects. Results are similar to Table 6, but less precise.
- Conclusion: Fascinating, though the evidence is not as conclusive as one would like.
- It is a virtue of this study that the quasi-experiment takes place in a natural (albeit unusual) market setting. That is, the sample is composed of real, high-stakes employment decisions. We don't have to worry about Hawthorne effects, demand effects, and other distortions potentially induced by a laboratory environment.
- Does this study provide evidence that orchestras engaged in taste-based discrimination *or* statistical discrimination? We cannot tell. It could be that females did worse in the non-blind condition due to taste discrimination. On the other hand, it could be the case that orchestras statistically discriminated against women in the non-blind condition. The blind condition rules out either taste-based or statistical discrimination—since neither can operate when the orchestra cannot determine the gender of the performer. So, the contrast here is between no discrimination (blind audition) versus unknown form(s) of discrimination (in the non-blind condition).

- Could these results also be explained by 'stereotype threat'—i.e., women performed better when auditions were blinded because they knew there was no discriminatory expectation? It would be fascinating to assess this. This would be testable if we had recordings of the auditions that could be evaluated by independent experts. Note that if women and/or men performed systematically worse or better under the blind versus non-blind condition, that would invalidate our interpretation of the findings above. Hence, our conclusions rest on the untested assumption that this is *not* occurring.

4.2 A Field Experiment in the Labor Market: Bertrand and Mullainathan (2003)

There have been many audit studies of hiring. In the canonical study, black and white applicants with identical resumes who are trained to behave similarly during interviews apply for the same jobs. The data analysis consists of testing whether the white applicant is more likely to receive a job offer than the black applicant.

There are numerous limitations of such studies:

1. Tiny samples b/c expensive
2. Not double-blind. Experimenters may have an agenda.
3. Artificial setting: Experimenters do not intend to complete the transaction (e.g., take the job, buy the car, rent the apartment...)
4. Could also be biased by 'stereotype threat'—perhaps the black applicants perform worse due to the presence of negative expectations.
5. Do not necessarily measure discrimination 'at the margin.'

Great idea: Apply for jobs by sending resume by mail or fax. Manipulate perceptions of race by using distinctively ethnic names. Otherwise, hold constant resume characteristics. Are 'callback' rates lower for distinctively black-named applicants? (B&M are not the first to do this, but they did it on a large scale and their work received considerable attention.)

Let's again briefly consider the statistical framework. Akin to the prior case where we were *not* manipulating gender, here we are not manipulating *race*. But we are manipulating percep-

tions of race by randomly assigning 'black-sounding' or 'white-sounding' names to otherwise identical resumes. Denote each resume by i. We write the probability of callback for resume i given a white-sounding (Y_0) or black-sounding (Y_1) name as:

$$\begin{aligned} E[Y_{0i}] &= \alpha_i, \\ E[Y_{1i}] &= \alpha_i + \gamma N_i. \end{aligned}$$

Hence, the estimated treatment effect in this case is:

$$\hat{T} = E[Y_{1i} - Y_{0i} | N_i = 1] = \gamma.$$

Notice that we do not need a 'second contrast' as we did in the blind versus non-blind audition comparison for males versus males; the only treatment in this study is the 'blind' condition. In the prior study, we were concerned that performers who chose to perform in blind versus non-blind auditions might not be not comparable to one another. We solved this problem by limiting the sample to performers who performed in both audition types. We don't have a similar concern in this case because we have randomized the assignment of names to resumes. So, this virtually guarantees comparability in a large sample: $E[\alpha | N = 1] = E[\alpha | N = 0]$.

More formally, think of our basic comparison of hiring odds of black and white-sounding resumes:

$$\hat{T} = E[Y_1 | N = 1] - E[Y_1 | N = 0] = (E[Y_1 | N = 1] - E[Y_0 | N = 1]) + (E[Y_0 | N = 1] - E[Y_0 | N = 0]).$$

In general, the first term following the equal sign is the contrast of interest (i.e., the difference in hiring odds for the same resume assigned a black and white sounding name) and the second term is a bias term (equal to the difference in potential outcomes for resumes with black and white sounding names). Because of the random assignment, however, we can be reasonably confident that this bias is close to zero ($E[Y_0 | N = 1] - E[Y_0 | N = 0] = 0$). Hence, the contrast in hiring odds between white-sounding and black-sounding resumes will be due to the random name assignment and no other factor.

Let's examine the results.

- Table 1: Short answer is yes. Callback rates are lower for black sounding names.

- Table 2: In most cases, names receive equal treatment (a criticism that Nobel laureate James Heckman has levied against audit studies more generally). But that's because in most cases, applicants are not called back.
- Table 6: Discrimination based on zip-code characteristics appears quite important and does not systematically differ between white and non-white names. (Authors also view this as evidence against statistical discrimination.)
- Table 8: Considerable overlap in distributions of outcomes between white and black sounding names

Conclusion: Controversial paper with striking findings. Will spark a great deal of other research in this vein. Questions remaining: How do we translate callbacks into outcomes we care about? Does this provide any information about 'discrimination on average vs. discrimination at the margin'? Does not sort out taste from statistical discrimination.

4.3 Stereotype Threat: Steele and Aronson (1995)

In this highly influential 1995 article, Steele and Aronson characterize Stereotype Threat (ST) as a "predicament" that African American males find themselves in due to "inferiority anxiety." This is a concise way of phrasing it, and also suggests why ST is not an economic model of discrimination. Because economic models are primarily about people making rational decisions, they do not readily accommodate the idea of people performing poorly because the stakes are high. One could add this idea to a formal economic model, of course, but it wouldn't necessarily provide much insight.

The paper presents several experiments that test the Stereotype Threat (ST) proposition. We'll discuss three of these, ignoring Study 1 (which is inconclusive).

4.3.1 **Study 2:**

Subject pool: 20 black and 20 white Stanford students.

"Participants who signed up for the experiment were contacted by telephone prior to their experimental participation and asked to provide their verbal and quantitative SAT scores, to

rate their enjoyment of verbally oriented classes, and to provide background information (e.g., year in school, major, etc.). When participants arrived at the laboratory, the experimenter (a White man) explained that for the next 30 min they would work on a set of verbal problems in a format identical to the SAT exam, and end by answering some questions about their experience."

> Participants in the *diagnostic* condition were told that the study was concerned with "various personal factors involved in performance on problems requiring reading and verbal reasoning abilities." They were further informed that after the test, feedback would be provided which "may be helpful to you by familiarizing you with some of your strengths and weaknesses" in verbal problem solving. As noted, participants in all conditions were told that they should not expect to get many items correct, and in the diagnostic condition, *this test difficulty was justified as a means of providing a "genuine test of your verbal abilities and limitations so that we might better understand the factors involved in both."* Participants were asked to give a strong effort in order to "help us in our analysis of your verbal ability."
>
> In the *non-diagnostic* condition, the description of the study made no reference to verbal ability. Instead, participants were told that the purpose of the research was to better understand the "psychological factors involved in solving verbal problems.. . ." These participants were also told that they would receive performance feedback, but it was justified as a means of familiarizing them "with the kinds of problems that appear on tests [they] may encounter in the future."

Key findings

In the Non-diagnostic condition, Blacks and Whites performed equally well on the SAT test.[1] In the diagnostic condition, Blacks performed worse than Whites. The magnitude of this effect is hard to gauge, but it is highly significant. Why do we say it's hard to gauge? Steele-Aronson reports that the mean SAT scores of Black participants was 603 and of Whites

[1] Actually, that's not quite true. The authors regressions adjust these scores for paricipants' SAT scores; only the regression-adjusted scores are tabulated and we don't know how the raw scores compare (though the text says that black scores were on average lower). This regression adjustment is in fact a *terrible* idea if the authors believe that *ST* threat affects blacks' performance on the SAT. If so, the SAT score is itself potentially contaminated by stereotype threat.

was 655. The comparison on the non-diagnostic portion of the experiment gives the number adjusted for the SAT score. This makes it uncertain what the actual number was for either group and how large the true score reduction for blacks would be in the diagnostic condition if the authors didn't adjust for SAT score.

4.3.2 Study 3:

To better understand the finding above, we'd like to have to some evidence that the relatively poor performance of blacks in the diagnostic condition is *directly* related to ST—that is, something about the treatment triggers race-related anxieties among black subjects. To test this, Steele-Aronson measure whether stereotypes and self-doubts are 'activated' for blacks under the diagnostic condition.

Instructions to participants:

> **Diagnostic:** Because we want an accurate measure of your ability in these domains, we want to ask you to try as hard as you can to perform well on these tasks. At the end of the study, we can give you feedback which may be helpful by pointing out your strengths and weaknesses.
>
> **Non-diagnostic:** Even though we are not evaluating your ability on these tasks, we want to ask you to try as hard as you can to perform well on these tasks. If you want to know more about your LAP and HVR performance, we can give you feedback at the end of the study.

According to several metrics (see figures), blacks appeared to made significantly more anxious than whites in the diagnostic condition and, more importantly, they showed greater evidence of 'stereotype activation' and 'stereotype avoidance' under the diagnostic treatment (the tests for these behaviors are explained in the article).

4.3.3 Study 4:

The final study is the most famous. While the prior experiments manipulated potential anxiety levels, these manipulations were not necessarily directly tied to race—the 'queues' for triggering anxieties were not race-based. This fourth study is similar to Study 2 above, except that in

the diagnostic ('race prime') condition subjects were asked to record their *race*, gender and age prior to taking the test whereas in the non-diagnostic ('no race prime') condition, subjects were only asked to record age and gender.

The contrast in test scores is really quite remarkable (though, as always, shrouded in some uncertainty due to the authors' data manipulations). This set of results suggests that (a) it does not take much to activate stereotype threat, and (b) its consequence for performance may be quite significant.

4.4 Distinguishing Statistical from Taste-Based Discrimination: List (2004)

This study is a bit more challenging to follow than the others. That may be because the objective of the study is more ambitious than the first two—to test whether discrimination occurs in a 'well-functioning' marketplace and, if so, to evaluate whether that discrimination is 'taste-based' or statistical (or a combination of both).

The setting of the study is a sportscard trading market. Sportscard trading is a popular avocation, and apparently one on which List spends a great deal of time (one guesses that this was true even prior to his research in this area).

In Part I of the experiment, List recruits volunteers at a sportscard show to buy (from dealers) and sell (to dealers) a 1989 *Upper Deck* Ken Griffy Jr. PSA graded "9" baseball card. Apparently, this is a valuable commodity; in the experiment, subjects typically paid over $100 for this card when *buying* from dealers. When *selling* the card to dealers, they typically received about $30 (so, there is quite a large buy-sell spread).

There are many details to the experiment that we will not summarize. The main results of the initial experiment are evident in Table II:

1. Initial offers made by dealers to minorities (nonwhites, females and men over age 60) for transacting on the Ken Griffy card are inferior to initial offers to white males. Inferior means *high asking price* when the subject is buying from the dealer and *low offer price* when the subject is selling to the dealer.

2. Discrimination appears much greater in the treatment in which subjects are selling the card to the dealer than in which they are buying (no theory for this).

3. Final offers to minorities are not *as inferior* as initial offers to minorities (relative to white males).

4. But, minorities spend more time bargaining to achieve similar results to white males—suggesting that they have to expend resources to overcome discrimination.

5. Experienced dealers discriminate *more* than inexperienced dealers (not visible in Table II).

Thus, discrimination in this market is *amply* evident. But what is the nature of this discrimination? List distinguishes three possibilities: animus-based discrimination; statistical discrimination; and differences in bargaining ability.

4.4.1 The dictator game

The dictator game is a widely-used laboratory experiment. In this particular version, List gives dealers envelopes with five $1 dollar bills and informs the dealer the race/gender of the person to whom he or she is randomly paired (white male age 20-30 WM, white female age 20-30 WF, nonwhite male age 20-30 NMM, white male age 60+ WMM). The dealer anonymously decides how much of the money to keep (take out of the envelope) and how much to leave for the anonymous partner whom s/he will never (knowingly) meet.

Figure I summarizes the results of this experiment. Dictators seem to favor white females. White females are significantly less likely to receive zero dollars and more likely to receive two or three dollars. There is no race differences (nonwhite females were not tested, presumably because there were almost no nonwhite females active in this market). Notice that the pattern of discrimination *favoring* females is opposite to the trading patterns on the floor where females receive worse offers uniformly.

One may object that this dictator game is highly artificial and so the behavior observed here may not match what a dealer would do in a less artificial setting. This particular experiment probably has limited external validity.

4.4.2 The 'Chamberlain' market

This is a subtle and complex experiment that is designed to evaluate whether dealers *believe* that minorities are more or less effective at negotiating than nonminorities, and whether this belief explains why dealers make less favorable offers to minorities.

To evaluate these hypotheses, List sets up a market in which dealers and nondealers trade for 'customized' (i.e., defaced) baseball cards (so they have no outside market value) using real money. Dealers and non-dealers are each randomly assigned a reservation value for the card (so, the dealer may be willing to pay no more than $20 and the seller may be willing to accept no less than $15). The participants bargain over the price and then each gets to keep his or her surplus (so, in the numerical example, there is $5 surplus to allocate; if the card sells for $19, the seller gets one dollar in surplus and the buyer gets $1 in surplus).

The key manipulation in this study is this: in some experiments, the dealers *are told* that the sellers' reservation values are randomly assigned; in others, they are *not told*.

There are many possible predictions for this experimental setting depending on the underlying model. Let's just focus on two.

1. In the animus-based case, we'd expect minorities to fare worse under both the 'informed' and 'uninformed' cases, since dealers should bargain harder with minorities due to animus.

2. In the case of statistical discrimination, we should expect dealers to bargain harder with minorities if the dealers *do not* know that reservation values are randomly assigned; if they do know that reservation values are randomly assigned, they should treat minority and nonminority sellers similarly. This is particularly true if dealers *knowingly* statistically discriminate, since they should consciously 'shut down' their discriminatory behavior when it is not rational to discriminate (because reservation values are randomly assigned).

Main results:

1. Majority buyers outperform minority buyers when dealers *do not* know that reservation values are determined randomly. This is consistent with either animus-based or statistical discrimination.

2. Majority buyers perform similarly to minority buyers when dealers *do* know that reservation values are determined randomly.

These results suggest that dealers discriminate only when they do not know that minority buyers have randomly assigned reservation values. This suggests statistical discrimination—dealers believe that minorities are willing to accept less money than nonminorities for similar items.

4.4.3 Reservation value experiments

We will not discuss these experiments in detail, but to summarize: the idea is to directly test whether (a) minorities have a different distribution of reservation values from non-minorities; and (b) dealers believe this to be true.

The nice set of findings here is that minority reservation values are more *dispersed* than non-minority reservation values, meaning that all else equal, minorities are more likely to have low reservation values (even if on average, their valuations are similar to nonminorities). (Figures II and III are pretty compelling.) This makes it rational for dealers to make them less favorable offers and to bargain harder with them.

A final experiment suggests that dealers know that these reservation value distributions differ. In particular, dealers—especially experienced dealers—are able to guess more accurately than chance would predict which distribution belongs to which race and gender group.

4.4.4 Conclusions of List study

This study offers a clever and rigorous effort to evaluate whether discrimination exists in the marketplace and from where it arises. The evidence offered here strongly suggests that *statistical discrimination* in the sportscard market is largely responsible for differential bargaining behavior of dealers facing minority versus non-minority buyers. The internal validity of these conclusions looks very solid. What about the external validity?

4.5 Further evidence on statistical discrimination: Fershtman and Gneezy (2001, QJE) [Will not be discussed in class.]

There are numerous experimental studies of discrimination. This one has particularly puzzling and provocative results. Basic idea of this study: Randomize the ethnic identity of the player faced in a series of games. Look for disparate treatment in games that attempt to isolate taste-based and statistical motives for possible discrimination.

Question: Do Ashkenazic and Eastern Jews appear to discriminate against one another. Four questions:

1. Is there differential treatment based on ethnic affiliation?
2. Does the discrimination reflect group bias in that each player favors members of his own group – or is there systematic discrimination by all against some?
3. Is discrimination simply taste-based or does it reflect the players' assessments of the differing reactions of members of groups to their actions (i.e., retaliation).
4. Are these assessments (stereotypes) accurate?

On this fourth point, note that statistical discrimination based on bad statistics (i.e., inaccurate assessments) is notionally distinct from simple taste-based discrimination, but unlike 'accurate statistical discrimination,' it's not a 'rational' economic response to group differences. Another word for inaccurate statistical discrimination is prejudice.

All experiments in this paper involve transfers of money between players. Transactions are done without personal contact, but names are written on paper, and are identifiably ethnic. All transactions occur over several days – so, the expectation of immediacy that might act like a disciplinary device for cooperation is absent. All results focus on males until end of paper.

Findings:

- **Trust game**:
- Transfer money to a player. Experimenter triples it. Recipients decides how much to return.

- Figure I: Male A Jews receive much larger transfers on average than male E Jews. In fact, 60% get the full amount (approximately equal to \$25 $\times$ 3).

- Figures IIIa and IIIb: Remarkably, E and A Jews treat E and A Jews similarly. In other words, neither group appears to 'trust' E Jews as much as A Jews.

- Is this statistical discrimination – i.e., are E Jews less 'trustworthy?' See Table II. There do not appear to be any differences in average amounts returned by E and A Jews (would be interesting to see distributions as well as means).

- **Taste for discrim: Dictator Game.**

- Same as trust game – transfer money, experimenter triples – but in this case, Player B is completely passive. No money can be returned. So, it's just taste for generosity.

- See Figure IV. Here, the average transfers to each ethnic group are similar but the distributions are not. More A Jews received zero, more E Jews received 5. Again, the ethnicity of the sender was not predictive of the amount sent – only the ethnicity of the recipient.

- **Stereotypes: Reaction to unfair treatment. Ultimatum game**

- There is a belief that E Jews are more likely than A Jews to act harshly in response to a perception of unfairness, i.e., have a greater sense of honor/humiliation.

- Test this with Ultimatum game. Here, Player A proposes a division. The amount given to Player B is tripled. Then Player B can accept or refuse.

- See Figure V. In this case, E Jews generally receive more. The modal A Jew receives 5 of 20 NIS (which will be tripled to 15, so this is an equal division ex post). The modal E Jew receives 10 of 20 NIS, which will be tripled to 30, so he gets a larger share of pie.

- One interpretation: Honor. A second interpretation (not in paper): Pandering. Players may not believe that E Jews will correctly understand that '5' is an equal division – and hence will reject.

- Is this discrimination rational? No evidence that E Jews are systematically more likely to reject an offer...
- **Gender differences:**
- Women do not appear to take ethnicity or gender into account.
- **Conclusions**
- A puzzling mix.
    1. There are clearly major differences in how males treat other males by ethnicity.
    2. But, this disparate treatment is not own-group biased. Both E and A Jews treat E Jews differently from A Jews. It appears they trust E Jews less.
    3. But E Jews do not appear less trustworthy conditional on receiving a transfer.
    4. Does not appear motivated by animus. In the Dictator game, groups receive roughly equal treatment.
    5. In the ultimatum game, behavior of proposers is consistent with beliefs that other group will conform to stereotype of honor/humiliation.
    6. But no evidence that groups do behave this way.
- Appears to be 'statistical discrimination based on bad statistics...'
- Question: How much can we learn from lab experiments about the importance of discrimination in markets at the margin (as opposed to on average). One can choose to draw strong inferences from these findings or dismiss them entirely because they are 'artificial.'