Racial Perceptions, Stereotype Threat, and the Black-White Test Score Gap: A Bayesian Principal Stratification Approach

Daniel E. Ho†
Stanford Law School
August 20, 2006

Abstract

Does stereotype threat explain part of the black-white test score gap? Recent evidence suggests that the perception of a white test administrator causes black test takers to perform worse, but that race of the test administrator itself has no impact independent of racial perception. If correct, this finding sheds new light on the validity, mechanism, and normative legal implications of stereotype threat. Using a Bayesian principal stratification approach – a framework that makes explicit and relaxes crucial assumptions implicit in conventional instrumental variables (IV) analyses – I reanalyze a quasi-experiment in a survey of political knowledge during which respondents only interact orally with the administrator. Examining the race of the interviewer alone, I find more evidence of a stereotype lift for whites than a threat for blacks. Moreover, racial effects do not appear to be mediated by racial perceptions alone. The administrator’s race affects performance similarly, regardless of the respondent’s perception of administrator’s race. This suggests that conscious racial perceptions are not crucial to the operation of stereotype threat. The evidence for stereotype threat from this experiment, however, is a mixed bag. Methodologically, the article illustrates widely-applicable methods for researchers to gain more leverage and assess the credibility and sensitivity of IV studies.

†Assistant Professor, Stanford Law School, 559 Nathan Abbott Way, Stanford, CA 94305. Email: dho@law.stanford.edu, URL: people.iq.harvard.edu/~dho
1 Introduction

The black-white test score gap remains a stark empirical regularity in standardized testing (see, e.g., Jencks and Phillips, 1998). Black students consistently perform worse than white students, scoring, for example, almost 200 points lower on the SAT between 1998-2001 (Card and Rothstein, 2005). Yet extant observational research has provided limited insights into the causes of the test score gap, in part because of the difficulty of drawing causal inferences when variables of interest (e.g., segregation, school quality, family structure, ability) are difficult to measure, confounded with characteristics that independently affect test scores, and themselves affected by race.

Fryer and Levitt (2004, 2006) examine the black-white test score gap in the first few years of school, finding, contrary to extant scholarship, that a relatively sparse covariate set can reduce the gap at the beginning of kindergarten. They conclude that school quality remains one of the only plausible explanations of the growth of the gap after kindergarten, finding no evidence for rival explanations such as parental, environmental, or neighborhood effects, bias in standardized testing, or discrimination. They “do not directly test other social-theoretic explanations such as stereotype threat” expecting that stereotype threat does not affect the young children in their sample (Fryer and Levitt, 2004, p. 460 n.26), and find no detectable effects of teacher race has on black student performance.

At the same time, experimental studies have uncovered considerable evidence for “stereotype threat,” a theory that posits that racial stereotypes of inferior intellect cause poor performance. The activation of a stereotype is hypothesized to induce anxiety in the black test takers of confirming the stereotype, thereby leading to poor performance. In a series of seminal experiments, Steele and Aronson (1995) found that black students underperformed relative to white students when told that a test was a diagnostic of intellectual ability, whereas the two groups performed indistinguishably when informed that the test was not diagnostic. Similarly, black students underperformed when merely asked about their race before the test. Since then the effects of activating stereotypes has been documented in a host of other settings (see, e.g., Hoff and Pandey 2006, Spencer, Steele and Quinn 1999). If true, it is argued, stereotype threat may explain part of the black-white test score gap (Helms 2005). The evidence, however, of the operation of stereotype threat outside of a
controlled laboratory so far is mixed (Sackett, Hardison and Cullen 2004; Carneiro, Heckman and Masterov 2005 (“No serious empirical scholar assigns any quantitative importance to stereotype threat effects as a major determinant of test score gaps.”)) – little evidence connects experimental and observational studies, and the precise dynamics somewhat unknown (Smith 2004). It remains unclear, in particular, whether racial effects, to the degree they exist, are mediated through perceptions of race or whether racial effects exist (if at all) independent of conscious racial perceptions.

The validity and mechanism of stereotype threat would have considerable impact on law and social science (see, e.g., Hanson and Yosifon 2004). Stereotype threat may, for example, account for outcomes differences between immigrants and American-born minorities (Guinier 2002), inform the use of non-legal sanctions for criminal behavior (Teichman 2005), and explain part of the difference in bar exam performance (Clydesdale 2004), thereby arguably informing research on the impact of affirmative action (see Sander 2004; Ho 2005b; Sander 2005; Ho 2005a; Ayres and Brooks 2005). Moreover, whether stereotype threat is induced through conscious racial perception or implicit actions bears large implications for antidiscrimination law (cf. Jolls and Sunstein 2006b; Lawrence III 1987). What degree of weight, for example, should courts accord to employee aptitude tests, when these tests themselves may induce racial bias? What steps should employers take to remove stereotype-inducing environmental cues? Should employers be able to proffer the fact that adverse action was taken without face-to-face interaction with the employee? And whether stereotypes are conscious may affect the success of normative prescriptions, such as consciously manipulating stereotypes to “debias” the effects of stereotype threat (Kray, Galinsky and Thompson 2002; Jolls and Sunstein 2006a). Isolating the mechanism of stereotype threat may in turn inform the design of standardized testing to minimize racial bias (Kang and Banaji 2006).

This paper focuses a telephone survey quasi-experiment to assess how racial perceptions affect respondent answers. Davis and Silver (2003) (DS) first analyzed this dataset examining primarily the causal effect of racial perceptions on answers to factual political knowledge questions. Stereotype threat implies two hypotheses. First, black respondents should perform worse on survey responses when perceiving a white interviewer: “the ‘threat’ [of negative intellectual capacity] is likely to be induced by the perception that the interviewers are white, not directly by whether the interviewers are actually white” (Davis and Silver 2003).
Second, white respondents should not be affected by the race, or perception of race, of interviewers. Accordingly, extant analyses of the data found that the perception of race, but not race itself, negatively affects answers by black respondents, and that race does not affect white respondents. If correct, these findings prove to illuminate the complex dynamics of race. Yet they are not uncontroversial. The chief findings appear at odds with implicit bias research, which suggests that conscious perception is not necessary for bias to manifest itself, as well as other studies finding that majority groups may in fact affected by stereotype threat (Danso and Esses, 2001; Aronson et al., 1999; Walton and Cohen, 2003).

This paper shows that the previous findings are artifacts of endogenous racial perceptions. Previous analyses of the data that simply regress outcomes on expressed perception of the interviewer’s race have ignored the most crucial part of the survey design: the race of the interviewer was in fact physically randomized. Capitalizing on this randomized instrument of the race of the interviewer to overcome endogenous racial perceptions, I find that the race of the interviewer is much more likely to affect white respondents, and that the effect is unlikely to be mediated by racial perceptions alone. Even respondents who incorrectly assessed the race of their interviewer are affected by the mere assignment, not perception, of race. These results are consistent with other findings on stereotype lift (Walton and Cohen, 2003), as well as general research in implicit bias, suggesting that conscious racial perception is not necessary for race to affect performance. Nonetheless, the findings are a mixed bag of evidence for stereotype threat: the race of the interviewer does not appear to have direct effects on blacks.

Methodologically, this paper illustrates a Bayesian principal stratification approach to analyze a quasi-experimental design (Imbens and Rubin, 1997a,b). As is well-known, and as demonstrated below, causal inference in observational studies may be notoriously difficult without (or ignoring) physical randomization of treatment. Yet even when the treatment of interest cannot be randomized, randomized instruments may enable researchers to draw causal inferences of theoretical interest (e.g., Angrist, 1990; Angrist and Lavy, 1999; Howell and Peterson, 2002; Gerber and Green, 2000). Such quasi-experimental designs thereby bear much promise in social science generally, where randomization of the treatment itself is often infeasible or unethical. A quasi-experimental design also enables researchers to test causal effects outside of the confines of a controlled laboratory, thereby potentially bridging the chasm between experimental findings
of stereotype threat and observational studies. The approach illustrated here permits researchers to test the sensitivity of inferences to crucial exclusion restrictions commonly imposed in instrumental variables (IV) studies \cite{hirano2000structure,frangakis2002principal,frangakis2002principal,angrist1996instrumental}, thereby more directly testing observable implications of stereotype threat.

The paper proceeds as follows. Section 2 describes the political knowledge data, and problems inherent in assessing causal effects of a possibly endogenous variable, such as racial perception. Section 3 discusses the framework of causal inference that capitalizes on the random assignment of an instrument (interviewer race) to draw causal inferences about the “treatment” of interest (interviewer race as perceived by respondent). Section 4 formalizes the Bayesian pattern-mixture model used to implement the framework for the political knowledge data. Section 5 discusses results and Section 6 concludes.

2 The Political Knowledge Data

The data stems from the Michigan State of the State survey conducted in 2001. The data includes $i = 1, \ldots, N$ telephone survey respondents, all from Michigan. Each respondent was randomly assigned the race of a phone interviewer, denoted by $Z_i = 1$ if the interviewer identified herself as black, and 0 if white, which I refer to as assignment. Only black and white interviewers participated in this part of the survey. The treatment of interest is respondent $i$’s perception of the race of the interviewer, denoted as $T_i = 1$ if the respondent perceived the interviewer to be nonwhite, and 0 if white\footnote{The nonwhite category includes respondents who answered “black”, “other”, “don’t know”, or refused to answer the question. To account for these other categories, compliance behavior could alternatively be modeled by a multinomial model. With a multinomial logit, by independence of irrelevant alternatives (IIA), the approach is equivalent to the conditional dichotomization used here, and one could easily examine other perception categories. A multinomial probit relaxes the IIA assumption, but as the number of respondents in each of the nonwhite and nonblack categories is small, there is little power to identify the variance-covariance matrix.}.

The survey also includes three pre-treatment covariates of gender, equal to 1 if the respondent is male and 0 if female, college degree, equal to 1 if the respondent graduated college and 0 otherwise, and college attendance, equal to 1 if the respondent attended some college but did not receive a degree and 0 otherwise. Let $X_i$ denote the row vector of these pre-treatment covariates for each respondent $i$. The outcomes of interests are answers to seven political knowledge questions denoted by $Y_{ij}$ for questions $j = 1, \ldots, 7$, which equals 1 if respondent $i$ answered the $j$th question correctly, and 0 otherwise. Table 1 presents the questions, answers, and proportions answered.
Table 1: Proportion of political knowledge question answered correctly by 212 black and 221 white respondents. The third column reports $p$-values from Fisher’s exact test.

<table>
<thead>
<tr>
<th>Political Knowledge Question</th>
<th>Answer</th>
<th>Proportion Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum voting age</td>
<td>18 years</td>
<td>White 0.85</td>
</tr>
<tr>
<td>Presidential term limits</td>
<td>2 terms</td>
<td>White 0.90</td>
</tr>
<tr>
<td>Majority party in state legislature</td>
<td>Republican</td>
<td>White 0.62</td>
</tr>
<tr>
<td>Vote required to override presidential veto</td>
<td>$\frac{2}{3}$ vote</td>
<td>White 0.42</td>
</tr>
<tr>
<td>Number of US Supreme Court Justices</td>
<td>9 Justices</td>
<td>White 0.32</td>
</tr>
<tr>
<td>Term length of US Senator</td>
<td>6 years</td>
<td>White 0.34</td>
</tr>
<tr>
<td>Office of William Rehnquist</td>
<td>Chief Justice</td>
<td>White 0.29</td>
</tr>
</tbody>
</table>

correctly by black and white respondents. The raw differences are consistent with the general descriptive findings on the black-white test score gap: for all but one question white respondents perform better, and the difference is statistically significant for 5 of the 7 questions at $\alpha = 0.05$. For example, 32% of white respondents know that there are 9 Justices, compared to 22% of black respondents ($p$-value=0.03).

To see whether we can infer any causal role to racial perceptions, Table 2 examines covariate balance (see Ho et al., 2004). While Table 2 suggests that there is relative balance for observed covariates along the perception of race, any inference using a non-randomized regressor such as racial perception relies on the assumption that conditional on observed covariates, the assignment of racial perceptions is random. In the econometric literature, this is termed as conditional exogeneity or selection on observables (Heckman and Robb, 1985), and in the statistical literature it is known as ignorability of treatment (Rosenbaum and Rubin, 1983). This may be violated for a variety of reasons, such as the nature of measuring racial perception. Most importantly, interviewers asked respondents at the end of the interview: “what do you think is my racial background?” Accurate responses are likely to be confounded with a host of other variables, such as social politeness, embarrassment, socioeconomic background (segregation), social awareness, and how seriously respondents take the survey. Indeed, more than three times as many white respondents answered “don’t know” to the question, indicating that actual responses as to race are not likely to be conditionally random.

The crucial design aspect of the survey that was ignored in previous analyses, however, is that we also observe race of the interviewer which is physically randomized and thereby ensures balance of observed and
<table>
<thead>
<tr>
<th></th>
<th>Interviewer race</th>
<th>Interviewer perceived</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>white (Zi = 0)</td>
<td>black (Zi = 1)</td>
<td>t-stat</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black respondents</td>
<td>(n=212)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interviewer black</td>
<td>0.68</td>
<td>0.00</td>
<td>1.00</td>
<td>0.37</td>
<td>0.85</td>
<td>−7.63</td>
</tr>
<tr>
<td>Perceived nonwhite</td>
<td>0.64</td>
<td>0.29</td>
<td>0.81</td>
<td>−7.90</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Questions correct</td>
<td>2.97</td>
<td>2.84</td>
<td>3.03</td>
<td>−0.87</td>
<td>2.75</td>
<td>3.10</td>
</tr>
<tr>
<td>Male (Xi1)</td>
<td>0.33</td>
<td>0.28</td>
<td>0.35</td>
<td>−1.00</td>
<td>0.33</td>
<td>0.32</td>
</tr>
<tr>
<td>College degree (Xi2)</td>
<td>0.20</td>
<td>0.15</td>
<td>0.23</td>
<td>−1.47</td>
<td>0.16</td>
<td>0.23</td>
</tr>
<tr>
<td>Some college (Xi2)</td>
<td>0.37</td>
<td>0.41</td>
<td>0.35</td>
<td>0.80</td>
<td>0.43</td>
<td>0.34</td>
</tr>
<tr>
<td>White respondents</td>
<td>(n=221)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interviewer black</td>
<td>0.56</td>
<td>0.00</td>
<td>1.00</td>
<td>0.23</td>
<td>0.79</td>
<td>−9.62</td>
</tr>
<tr>
<td>Perceived nonwhite</td>
<td>0.59</td>
<td>0.29</td>
<td>0.83</td>
<td>−9.46</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Questions correct</td>
<td>3.74</td>
<td>3.68</td>
<td>3.79</td>
<td>−0.49</td>
<td>3.81</td>
<td>3.69</td>
</tr>
<tr>
<td>Male (Xi1)</td>
<td>0.43</td>
<td>0.41</td>
<td>0.44</td>
<td>−0.34</td>
<td>0.48</td>
<td>0.39</td>
</tr>
<tr>
<td>College degree (Xi2)</td>
<td>0.37</td>
<td>0.37</td>
<td>0.36</td>
<td>0.13</td>
<td>0.36</td>
<td>0.37</td>
</tr>
<tr>
<td>Some college (Xi2)</td>
<td>0.28</td>
<td>0.30</td>
<td>0.26</td>
<td>0.67</td>
<td>0.27</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Table 2: Summary statistics of political knowledge survey.

unobserved covariates. This instrument permits us to explicitly test the inferences about racial perceptions, in a way that is robust to these problems of endogenous perceptions.

3 Framework of Causal Inference

To leverage randomized interviewer assignment, I employ the framework of modeling “compliance” by Imbens and Rubin (1997a). Compliance is a term coined from deviations of classical randomized experiments, where experimental subjects may not comply with treatment assignment. By extension, in IV studies (also termed “encouragement designs”), it refers to treatment status in response to the instrument: the key idea is that when a researcher cannot randomly assign treatment she may be able to assign an instrument that induces compliers to take the treatment. For example, McDonald, Hiu and Tierney (1992) studied the effect of flu shots on morbidity using randomized letters sent to doctors encouraging them to inoculate patients for the flu. “Compliers” are patients who receive the flu shot solely because of the randomized letter, and an IV estimate is the causal effect of the flu shot on this subpopulation of compliers. Similarly, Angrist and Lavy (1999) capitalized on the instrument of Maimonides rule, limiting class sizes in Israeli public schools to 40 students, to examine the effect of class size on test score outcomes – compliance would refer to the
difference in class sizes due to (effectively-random) class-splitting by Maimonides rule.

Here, compliance behavior is the respondent perception of race in response to assignment of interviewer race. Let \( D_i(z) \) represent an indicator of perception of race as a function of assignment. \( D_i(1) \) equals 1 if \( i \) would perceive the black interviewer as nonwhite, and 0 if \( i \) would perceive the interviewer as white. \( D_i(0) \) equals 1 if \( i \) would perceive the white interviewer as nonwhite, and 0 if \( i \) would perceive the interviewer as white. Four types of compliance behavior are thereby defined by \( C_i \):

\[
C_i = \begin{cases} 
  c & \text{if } D_i(z) = z \text{ (Compliers, or “racially-astute” respondents)} \\
  n & \text{if } D_i(z) = 0 \text{ (Never-takers, or “white-perceivers”)} \\
  a & \text{if } D_i(z) = 1 \text{ (Always-takers, or “nonwhite-perceivers”)} \\
  d & \text{if } D_i(z) = 1 - z \text{ (Defiers)} 
\end{cases}
\]

for \( z = 0, 1 \). The terms for the subtypes (compliers, never-takers, always-takers, and defiers) stem from encouragement designs in the biomedical literature; more appropriate labels in this application for the first three may be “racially-astute” (i.e., those who correctly perceive the race of the interviewer in response to interviewer assignment), “white-perceiver” (i.e., those who always perceive the interviewer to be white), and “nonwhite-perceiver” (i.e., those who always perceive the interviewer to be nonwhite) respondents, respectively. The remainder of the paper will use these terms (complier / astute, never-taker / white-perceiver, always-taker / nonwhite-perceiver) interchangeably.

Note that compliance behavior, i.e., the type of respondent, is not directly observable, but causal effects are only properly defined within each type \( \text{[Frangakis and Rubin 2002]} \). An “as-treated” analysis fails to account for the fact that compliers, the subpopulation of respondents who correctly change their racial perception due to assignment of interviewer, may be substantially different from noncompliers (i.e., white- and nonwhite-perceivers, and defiers) in observed and unobserved factors. The instrument of interviewer race circumvents this selection bias between respondents, since perception for this subgroup differs only as a result of assignment, which is random.

We can now also define the potential outcomes \( Y_i(z,D_i(z)) \) under each assignment, suppressing \( j \) for notational simplicity. Counterfactual responses are treated as missing data to be imputed. The overall
intention-to-treat effect (ITT) is then defined as a weighted average of ITT$_t$ effects for each subtype:

$$\text{ITT} = \sum_{t \in \{c, n, a, d\}} \text{ITT}_t \ast \frac{N_t}{N}$$

$$\text{ITT}_t = \sum_{i \in \{i: C_i = t\}} \left[ Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \right] / N_t$$

where $t \in \{c, n, a, d\}$, $N_t$ denotes the number of respondents in each subtype.

The identification assumptions usually imposed in IV studies to estimate ITT$_c$ are as follows:

**Assumption 1 (No Interference Among Units)** $Y_i(z, D_i(z)) \perp \perp Z_i$ for all $z = 0, 1$ where $i' \neq i$.

where $\perp \perp$ denotes independence. The independence assumption implies that the potential responses to survey questions by one respondent are not affected by the assignment of interviewer race of another respondent. Where the assignment of one might affect outcomes of others it is violated. The assignment of municipality services as an instrument for voter information, for example, may affect voters outside of the assigned municipality if voters move to different municipalities as a result of assignment (see, e.g., Lassen 2003). Here, the assumption might be violated if the prior racial composition of interviewees affects future interview behavior.

**Assumption 2 (Random Assignment)** $Y_i(z, D_i(z)), D_i(z) \perp \perp Z_i$ for all $i$ and $z = 0, 1$.

Random assignment of interviewer race is plausibly met by virtue of the standard random digit dialing of the survey: “[a]ssignment [of respondents] was not based on either the characteristics of the interviewer or the characteristics of the respondents” (Davis and Silver 2003, p. 37). Accordingly, Table 2 provides no evidence of incomplete randomization within substrata of respondent races. Nonetheless, it remains curious that the proportion of self-identified black interviewers differs appreciably across black and white respondents. While roughly 67% of interviewers of black respondents identified themselves as black, only 56% of interviewers of white respondents identified themselves as black ($t$-stat=2.5). This may be due to (a) clustering of 73 interviewers, which would bias standard errors downwards, or (b) undocumented effects arising from the regional stratification in the survey designed to ensure roughly equal number of black and white respondents.
To account for this imbalance, the analysis stratifies on respondent race, relaxing the randomization assumption to conditional exogeneity. To the degree that there are further unspecified errors in randomization of interviewer race with respect to unobserved covariates, causal inferences may be substantially more difficult. One natural extension to account for interviewer-specific effects would be a multilevel model (Gelman and Hill 2006).

**Assumption 3 (Monotonicity)** $D_i(1) \geq D_i(0)$ for all $i$.

Monotonicity rules out respondents who would perceive black interviewers to be white, and white interviewers to be nonwhite (i.e., no defiers). While plausible here, where there is no evidence of hostile respondent behavior which might imply the existence of defiers, it may not be in other settings. Take the case of randomized direct mailing of campaign advertisement as instruments for voter information (Gerber and Green 2000). Defiers might be less inclined to inform themselves about the campaign as a result of reading the direct mailing, leading classic IV estimates to be biased.

**Assumption 4 (Exclusion Restrictions on Subtypes)**

(A) *No effect of race on nonwhite-perceivers:* $P(Y_i(1,D_i(1)) = 1|X_i,C_i = n) = P(Y_i(0,D_i(0)) = 1|X_i,C_i = n)$

(B) *No effect of race on white-perceivers:* $P(Y_i(1,D_i(1)) = 1|X_i,C_i = a) = P(Y_i(0,D_i(0)) = 1|X_i,C_i = a)$

Exclusion restrictions (A) and (B) rule out any effect of race on those respondents for whom the race of the interviewer has no effect on racial perception. This formalizes the hypothesis that race should affect responses only via the *perception* of race. The use of non-randomized instruments often falter on this ground, as Donohue III and Wolfers (2006), for example, discuss in the context of instruments for capital punishment. More generally, these restrictions in conjunction with monotonicity hold that the instrument affects outcomes only in the subpopulation of compliers who are treated as a result of assignment.

The particular virtues of a Bayesian framework as applied to IV designs are the ability to (a) relax monotonicity to allow for defiers, and (b) relax exclusion restrictions to test whether race affects respondents
whose racial perceptions are not altered by assignment. Why might ITT effects occur in these subpopulations who are unaffected by interviewer race in their racial perceptions? Theories emphasizing linguistic differences in interviewers might predict large ITT effects for respondents who do not perceive racial cues, while theories emphasizing social desirability with particular social groups would predict no ITT effects on these subpopulations. In addition, the assignment of a black interviewer may induce informal changes in the administration of the survey script because of interviewer perceptions irrespective of respondent perceptions. Whether race itself or race mediated by racial perceptions affects responses is a question of potentially wide interest, which may be explicitly tested in this framework.

Of course, ITT effects for non-compliers might alternatively indicate that the method of assessing racial perception is simply unreliable, due to self-censoring when respondents do in fact perceive racial cues. ITT effects for non-compliers thereby indicate that either the theory or the measurement is violated in some fashion. Both are of direct interest to research in stereotype threat and the black-white test score gap.

4 A Mixture Model

I follow Hirano et al. (2000) and model the outcomes as a mixture of three types. Let \( \theta \) denote the vector containing all parameters. Each type of respondent has an outcome distribution that is modeled as a logistic regression:

\[
P(Y_i|C_i = t, Z_i = z, X_i, \theta) = \pi_{tz} = \frac{\exp(\alpha_{tz} + X_i\beta)}{1 + \exp(\alpha_{tz} + X_i\beta)}
\]

where \( \alpha_{tz} \) represents the intercept for each type \( t \in \{c,n,a\} \) and assignment \( z = 0, 1 \), and \( \beta \) represents the vector of three slope parameters for pre-treatment covariates. I assume equal slopes across subtypes and \( z \).

Relaxing this assumption is straightforward. The distribution of types is modeled as a multinomial logit:

\[
P(C_i = t | X_i, \theta) = \Psi_{ti} = \frac{\exp(X_i\psi_t)}{\sum_{k \in \{c,n,a\}} \exp(X_i\psi_k)}
\]

where for identification, each element of \( \psi_n \) is restricted to 0, and \( \psi_c \) and \( \psi_a \) are vectors of four parameters corresponding to covariates and an intercept. As a result, there are 17 parameters to be estimated in \( \theta = \{\alpha_{c1}, \alpha_{c0}, \alpha_{a1}, \alpha_{a0}, \alpha_{n1}, \alpha_{n0}, \beta, \psi_a, \psi_c\} \) under no exclusion restrictions. Exclusion restriction (A) implies that
\[ \alpha_{a1} = \alpha_{a0}, \] and (B) implies that \( \alpha_{a1} = \alpha_{a0} \). We can now write the complete data likelihood as:

\[
L(\theta | Z, D, Y, X, C) = \prod_{i \in \{ i : C_i = c \}} \Psi_{ci}(\pi_{czi})^{Y_i}(1 - \pi_{czi})^{1 - Y_i} \times \prod_{i \in \{ i : C_i = n \}} \Psi_{ni}(\pi_{nzi})^{Y_i}(1 - \pi_{nzi})^{1 - Y_i} \times \prod_{i \in \{ i : C_i = a \}} \Psi_{ai}(\pi_{azi})^{Y_i}(1 - \pi_{azi})^{1 - Y_i}
\]

I use the following conjugate prior, which adds fractional observations to the data to ensure proper form of the posterior (Clogg et al., 1991):

\[
p(\theta) \propto N \prod_{i=1}^{N} \prod_{t \in \{ c, n, a \}} \prod_{z=0}^{1} \prod_{y=0}^{1} \left\{ \Psi_{ti}(\pi_{tzi})^{y}(1 - \pi_{tzi})^{1 - y} \right\}^{2.5/N}
\]

Similar to Jeffrey’s prior in a binomial model, this prior shrinks the likelihood towards an equal number of subtypes.

5 Results

At the outset, I substantially replicated the main findings of DS with the original data (using the endogenous perception of race, and ignoring random assignment of interviewers). While the original DS analysis uses OLS to model the total number of correct answers, the analysis here uses a more appropriate logit model for binary responses to each of the questions.\(^2\) Below I first present results from an ITT analysis on the effect of interviewer race on the survey answers to ascertain what part of the racial gap may be explained by stereotype threat. The results suggest that black interviewers cause white respondents to perform better, but have no impact on black respondents. This is consistent with a notion of stereotype lift. I then present IV results on the effects of racial perception, with a particular focus on sensitivity to exclusion restrictions. These results suggest that racial effects are not mediated by perception, indicating that there exists some form of implicit bias independent of expressed racial perceptions.

5.1 Intention-to-Treat Analysis: The Effect of Interviewer Race

To estimate the impact of the race of the interviewer I use a logistic model for each political knowledge questions, separately for black and white respondents. Each model includes pretreatment covariates and

\(^2\) Assuming equiprobability, the sum of correct answers could be modeled as a binomial distribution, but even under that implausible assumption, OLS would be inappropriate. If a single analysis model is desired, an alternative may be an IRT-type model that jointly estimates latent abilities of respondents (rather than the sum of correct answers).
interviewer race. 10,000 draws of the posterior distribution are sampled with the Metropolis algorithm, using a flat prior and the multivariate normal as the jumping distribution, scaled by the asymptotic variance-covariance matrix, with maximum likelihood estimates as starting values. Table 3 presents summary statistics of the ITT effects for each of the questions. The results are surprising in light of stereotype threat theory. First, race of the interviewer appears to cause an (statistically significant) effect in the direction hypothesized for only one of the seven questions for black respondents: a black interviewer appears to increase the probability that a black respondent correctly notes the office of Justice Rehnquist by 7%. Second, even more surprising, interviewer race appears to have greater effects on white respondents than on black respondents. In three of the seven questions posed to whites, the posterior probability of a positive effect is greater than 97%, while the effect is negative and significant for one model (the number of Justices question). For example, a black interviewer increases the probability of a white respondent answering the Senate term question correctly by roughly 6%. Lastly, the effects do not appear to be correlated with the test score gap, as reported in column 3. While these results are not anticipated by DS, which hypothesizes that white respondents should remain unaffected by interviewer race, they are consistent with other studies finding that white respondents fare better with black experimenters (e.g., Danso and Esses, 2001). White respondents appear to experience a stereotype lift with black interviewers. While that lift may account for some of the test score gap, it cannot account for all of it.

<table>
<thead>
<tr>
<th></th>
<th>White Effect</th>
<th>S.D.</th>
<th>Black Effect</th>
<th>S.D.</th>
<th>Raw Test Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting age</td>
<td>0.13</td>
<td>0.03</td>
<td>0.01</td>
<td>0.03</td>
<td>-0.02</td>
</tr>
<tr>
<td>Pres. term</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>Maj. party</td>
<td>-0.02</td>
<td>0.03</td>
<td>-0.07</td>
<td>0.04</td>
<td>0.13</td>
</tr>
<tr>
<td>Veto override</td>
<td>0.08</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.21</td>
</tr>
<tr>
<td>No. Justices</td>
<td>-0.09</td>
<td>0.03</td>
<td>-0.03</td>
<td>0.03</td>
<td>0.10</td>
</tr>
<tr>
<td>Senate term</td>
<td>0.06</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.15</td>
</tr>
<tr>
<td>Rehnquist</td>
<td>-0.00</td>
<td>0.03</td>
<td>0.07</td>
<td>0.03</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Table 3: ITT results of the effect of a black interviewer. Raw Test Gap indicates the difference in proportion of correct answers.
5.2 Principal Stratification: The Role of Racial Perceptions

This section now turns to the role of racial perceptions. Separate mixture models were estimated for each question for black and white respondents, and I consider four combinations of exclusion restrictions: (a) no effects on white- and nonwhite-perceivers (akin to classical IV estimates), (b) no effects on white-perceivers, (c) no effects on nonwhite-perceivers, and (d) no restrictions at all.

The parameters are estimated via Markov Chain Monte Carlo, with a Gibbs sampler iterating between (a) drawing the latent subtypes and (b) drawing parameters conditional on the subtypes. Step (a) is simplified by the monotonicity assumption, which implies that $P(C_i = a | Z_i = 0, D_i = 1) = 1$ and $P(C_i = n | Z_i = 1, D_i = 0) = 1$ (i.e., no inastute respondents who always perceive race incorrectly). Hence, I draw from the mixture of compliers (astute respondents) and always-takers (nonwhite-perceivers) when $Z_i = 1, D_i = 1$ and mixture of compliers (astute respondents) and never-takers (white-perceivers) when $Z_i = 0, D_i = 0$. For efficiency, draws in step (b) were undertaken in sequential Metropolis steps for first the multinomial parameters \{\psi_a, \psi_c\}, second the subtype intercepts \{\alpha_{c1}, \alpha_{c0}, \alpha_{a1}, \alpha_{a0}, \alpha_{n1}, \alpha_{n0}\}, and finally the covariate slopes \{\beta\}, using a $t$-distribution with five degrees of freedom scaled by tuning constants as the jumping distribution. Tuning constants were chosen to minimize autocorrelation of the posterior draws. Four overdispersed chains of 14,500 draws each were run for each model, discarding the first 2,000 burn-in draws, saving every fifth draw, and combining the four chains for a posterior sample of 10,000 draws. Convergence was monitored by examining autocorrelation of all parameters and $\hat{R}$ statistics for overdispersed chains (Gelman and Rubin [1992]). The sampling algorithm was implemented in C++, using the Scythe matrix library (Martin, Quinn and Pemstein [2002]), and convergence diagnostics were conducted in R.

Table 4 presents quantities of interest from the posterior distribution of models imposing exclusion restrictions for white- and nonwhite-perceivers and no exclusion restrictions. For space limitations, and since results are materially the same, results are not reported for models relaxing exclusion restrictions (A) and (B) separately. Consider first the effects estimated on white respondents (the top part of Table 4), for whom it is hypothesized that race should not affect responses. Imposing both exclusion restrictions, the estimated causal effect of the perception of a nonwhite interviewer on racially-astute white respondents is a 25% increase in the probability of answering the voting age question correctly (one-tailed posterior $p$-value $< 0.1$).
Table 4: Summary statistics of posterior distributions of models with exclusion restrictions for white- and nonwhite-perceivers and no exclusion restrictions.
Figure 1: Simulation scatter plot of the joint posterior distribution of $ITT_n$ and $ITT_c$ in the models of white respondents (top panel) and black respondents (bottom panel) for the minimum voting age question with no exclusion restrictions.
Yet relaxing the exclusion restrictions demonstrates that white-perceivers exhibit substantial \( ITT_n \) effects as well: even for individuals who always perceive the interviewer to be white, nonwhite interviewers cause a roughly 18% increase in the probability of answering the question correctly – an increase that is statistically indistinguishable from \( ITT_c \) \((P(ITT_c \geq ITT_n) = 0.49)\).

The top panel of Figure 1 plots simulation scatter plots of the joint posterior distributions of \( ITT_n \) and \( ITT_c \) for white respondents to the voting age question. Most of the posterior mass for both white-perceivers and racially-astute respondents is concentrated above the origin, and the effect appears to be larger for white-perceivers, albeit with higher variance. Due to the relatively small sample size we unfortunately cannot estimate the parameters with a high degree of precision, particularly given the small proportion of noncompliers. Nevertheless, this demonstrates the strength of the Bayesian principal stratification approach: imposing exclusion restrictions that are inherent in most IV studies can lead us to draw overconfident inferences on compliers. Classical IV estimates are inflated because all race of the interviewer effects are attributed to perceived race, but the results suggest that race does not appear to affect responses exclusively via racial perceptions.

On the other hand, results from the number of Justices question as posed to white respondents indicate that the negative \( ITT \) effect is driven by expressed perceptions. Astute white respondents are 29% less likely to answer the question correctly \((p\text{-value}=0.04)\), an effect statistically significantly different from that on white-perceivers \((p\text{-value}=0.06)\). For one question then, effects do appear to be mediated by racial perceptions: whites sometimes do worse when they perceive a black interviewer; but overall, there still appears to be a stereotype lift not driven by conscious perception. Without knowing the question, white respondents have a slight tendency to perform better with black interviewers.

Consider now the effects estimated on black respondents. The only substantive complier effect when imposing all exclusion restrictions, namely for the majority party question, is in the opposite direction anticipated by stereotype threat theory. The effect of perceiving the interviewer to be nonwhite is roughly a 24% decrease in getting the answer right, but standard errors are large. And after relaxing the exclusion restrictions, the only significantly substantive effect of black interviewers in the posited direction is not on racially-astute respondents but on white-perceivers for the voting age question. The bottom panel of Figure 1...
shows that while the $ITT_c$ effect for that question is centered around the origin, over 90% of the posterior mass of $ITT_n$ is positive. White-perceiver blacks appear to respond more positively to black interviewers than astute respondents – the posterior probability that $ITT_n \geq ITT_c$ is roughly 86%. This contradicts the hypothesis that the effect should be observable only for astute respondents. Overall, however, there is little evidence that white interviewers, or the perception of such, affect black respondents at all.

To summarize the main findings across the 7 questions for blacks and whites, Figure 2 plots the posterior densities for each of the subtypes. Black solid lines denote the effect of a black interviewer on astute respondents ($ITT_c$); grey lines denote the effect of a black interviewer on nonwhite-perceivers ($ITT_a$), and black dashed lines denote the effect on white-perceivers ($ITT_n$). While some differences between the $ITT$ effects on subtypes appear to exist – compare, for example, nonwhite-perceiver effects for the voting age question, or the majority party effects for black respondents – generally the densities overlap substantially. On the other hand, noncomplier effects are centered around the origin for a number of questions, giving some credence to the exclusion restrictions.
The left panel of Figure 3 plots the 14 sorted one-tailed \( p \)-values for the overall ITT effect. If there are no effects, we would expect these values to be roughly uniformly distributed along the 45-degree line. We observe a slight kink towards the origin, suggesting systematic ITT effect (although a rank sum test yields a \( p \)-value of 0.14, so that the deviation may be due to sampling variability). Contrast that now with the middle and right panels, which present \( p \)-values of whether the ITT\(_e\) effect is greater than the ITT\(_a\) effect – more concretely, whether the race of the interviewer affects only astute respondents that correctly perceive race. These \( p \)-values, in contrast to the left panel, appear roughly randomly distributed. In sum, there appears to be some evidence of the effect of race of the interviewer, but the effect is not confined to astute respondents. Conscious perception does not appear necessary for ITT effects to manifest themselves.

## 6 Conclusion

The results are intriguing from a substantive perspective, although inconsistent the DS findings, which are likely to be an artifact of endogenous racial perceptions. Instead, consistent with (Aronson et al., 1999) and (Danso and Esses, 2001), white respondents appear susceptible to the race of the interviewer, exhibiting a stereotype lift. Moreover, these effects do not appear to be mediated by measured racial perceptions alone: the race of the interviewer may affect respondents in subtle ways of which respondents are not consciously aware. The results hence give some credence to shifting antidiscrimination law’s focus from overt to implicit bias (Jolls and Sunstein, 2006b). On the other hand, the quasi-experiment does not provide conclusive evidence of the operation of stereotype threat: black respondents do not appear to respond substantially to the assignment or perception of a white interviewer.

Methodologically, this paper has illustrated an application of an exciting new area of applied research, spawned by reinterpretations of traditional IV estimation strategies in a potential outcomes framework. The analysis shows that researchers can leverage randomized instruments to test hypotheses with greater scientific credibility and illustrates how to assess sensitivity of classical estimates to implicit assumptions.
References


21


