Bias in white: A longitudinal natural experiment measuring changes in discrimination*

RUNNING HEAD: Bias in white: Measuring changes in discrimination

Brian Rubineau
Cornell University
ILR School
394 Ives Hall
Ithaca, NY 14853
brubineau@cornell.edu
phone: 607-255-3048
fax: 810-963-2738

Yoon Kang
Cornell University

June 15, 2011

Word count: 10,127

***** We are grateful to Carol Storey-Johnson, Michael Slade, Anne Connolly, Latasha Boston, Sangchan Park, Evan Polman, Alan Moses, and Lynne Vincent for assistance in making this study possible; to the members of Cornell's CITRA seminar, Carla Boutin-Foster and Marty Wells for discussions and feedback as the research was being planned and conducted; and to Marya Besharov, Diane Burton, Emilio Castilla, Glen Dowell, Roberto Fernandez, Olga Khessina, Jason Schnittker, Tony Simons, Ezra Zuckerman, Robin Ely, Jesper Sørensen and participants of the ASQ Gender and Race in Organizations Workshop. We are also grateful to the Cornell University Institute for Social Sciences for a generous grant supporting this study. All errors are the authors' sole responsibility.

Bias in white: A longitudinal natural experiment measuring changes in discrimination

ABSTRACT *(145 Words)*

Many professions are plagued by disparities in service delivery. Racial disparities in policing, mortgage lending, and health care are some notable examples. Because disparities can result from a myriad of mechanisms, crafting effective disparity mitigation policies requires knowing which mechanisms are active and which are not. In this study we can distinguish whether one mechanism – statistical discrimination – is a primary explanation for racial disparities in physicians' treatment of patients. In a longitudinal natural experiment using repeated quasi-audit studies of medical students, we test for within-cohort *changes* in disparities from medical student behaviors as they interact with white and black patient-actors. We find significant *increases* in medical students' disparate behaviors by patient race between their first and second years of medical school. This finding is inconsistent with statistical discrimination predictions and challenges the idea that statistical discrimination is primarily responsible for racial disparities in patient care.

Bias in white: A longitudinal natural experiment measuring changes in discrimination

INTRODUCTION

For a wide range of occupational and professional roles, research has documented consistent and significant disparities arising from professionals' interactions with their clients. Examples of include racial profiling by police officers (Knowles et al. 2001), red-lining trends by real estate brokers (Yinger 1996) and mortgage lenders (Ladd 1998), foul calls by basketball referees (Price and Wolfers 2007), negotiated car sales prices (Ayres and Siegelman 1995), and critically for this study, racial disparities in the patient care delivered by physicians (Institute of Medicine [IoM] 2003). Once such disparities are identified, professions may work towards their mitigation. However, without an understanding of the mechanisms giving rise to disparities, professions are unlikely to design effective intervention strategies.

A diverse set of theorized mechanisms may all contribute to disparities, and empirically disentangling the active mechanisms from the inert is a difficult and challenging area of active research (e.g., Altonji and Pierret 2001; Chandra and Staiger 2010). These research efforts are crucial for informing effective disparity-reducing policies. Using a uniquely suited longitudinal natural experiment in the form of a repeated quasi-audit study of medical students, this paper reports on the elimination of one theorized mechanism – statistical discrimination – as a primary explanation for a very consequential disparity– racial disparities in physicians' treatment of patients.

The natural experiment of this study comes in the form of medical students "treating" race-varying *standardized patients* (SPs) – actors trained to portray a specific medical case. This common pedagogical practice yields a quasi-audit study. Medical student cohorts participate in repeated SP case encounters during their medical school training, generating longitudinal panel data of these quasi-audits. Audit studies are one of the best ways to measure disparities arising from discriminatory decision-making (National Research Council 2004; Quillian 2006), but until now, these studies have been entirely cross-sectional. This longitudinal study allows greater elucidation of the generative mechanisms for racial disparities resulting from professionals' behavior than available via the previous cross-sectional approaches.

Statistical Discrimination

The theory of statistical discrimination was originally put forth as an economic explanation for enduring disparities within labor markets (Arrow 1972; Phelps 1972). This theory helped to explain how disparate outcomes could endure in a market of rational actors, where previous theory suggested that disparities from discrimination should be competed away (Becker 1971). The appeal of a theory of discrimination based in rational behavior rather than bias may help to explain why statistical discrimination has often been adopted as an explanation for disparate outcomes in contexts beyond the

labor market. Statistical discrimination has been enlisted to explain racial disparities in outcomes from policing (Knowles et al. 2001), housing (Ross and Turner 2005), mortgage lending (Ladd 1998), customer service (Lee 2000), automobile markets (Ayres and Siegelman 1995) and health care (Balsa and McGuire 2001). These cited examples invoke statistical discrimination to explain racial disparities generated by the behavior of workers or professionals while acting in their occupational or professional roles.

Despite the compelling nature of the statistical discrimination explanation for the endurance of many observed societal disparities, little positive evidence supports this explanation (Correll and Benard 2006). The dearth of positive evidence for statistical discrimination may result from its relative unimportance among the many mechanisms contributing to disparate outcomes (see NRC 2004 and Pager and Shepherd 2008 for reviews of the multiple mechanisms underlying racial disparities), or simply result from the difficulty involved in empirically disentangling statistical discrimination from these other mechanisms. This difficulty is hardly surprising, given that different theorized mechanisms of discrimination are different attempts at explaining the same observable phenomena.

In the well-documented case of racial disparities in patient treatment by physicians, statistical discrimination is seen as a "potent source" (McGuire et al. 2008:2) for those disparities. Several studies of disparate care document findings consistent with statistical discrimination explanations (e.g., Lutfey and Ketcham 2005; McGuire et al. 2008). The research designs of these studies however, are not able to distinguish statistical discrimination from other mechanisms, and their findings are also consistent with other disparity-generating mechanisms such as prejudice (e.g., Fennell 2005:1714). This study attempts to falsify the statistical discrimination explanation for racial disparities in patient care by physicians. To be clear, this falsification cannot and does not show that statistical discrimination never contributes to disparities in care. Rather, we show that statistical discrimination is unlikely to be either the sole or primary mechanism responsible for racial disparities in patient care.

Our approach towards falsifying the statistical discrimination explanation for racial disparities in care is logically akin to the approach of someone who wishes to falsify the theory that human babies come from storks. The falsification of stork theory, and thus a demonstration of the existence and importance of other baby-generating mechanisms, can be accomplished by empirically documenting a context where there are no storks but where an increase in new human babies is nonetheless observed. We examine a setting where statistical discrimination would predict either static or decreasing levels of racial disparities, and find instead significant increases in racial disparities. This finding, disconsonant with statistical discrimination, shows the existence and importance of other mechanisms for generating racial disparities in patient care.

Defining Statistical Discrimination

Before attempting to falsify the statistical discrimination explanation for racial disparities in care, a clear definition of the mechanism is needed. If a particular important characteristic (e.g., productivity) is both hard-to-observe directly and has different distributions (i.e., in the means or variances) across more easily-observed social categories within a population (e.g., age), it may be rational to prefer to treat (e.g., hire) population members differently based on these more easily-observed categories (see Correll and Benard 2006; England 1994:60-63 for reviews). When a decision-maker makes decisions resulting in disparate outcomes by social category based on the true distributional differences associated with category membership, that decision-maker can be said to be engaging in statistical discrimination (Aigner and Cain 1977; Baumle and Fossett 2005: 1251; NRC 2004: 61-62).

In contrast, decisions based on prejudice, or biased or inaccurate perceptions of differences in the hard-to-observe characteristic by social category are not statistical discrimination. Discriminatory behavior that is based on erroneous perceptions is indistinguishable from and definitionally equivalent to discriminatory behavior from unfounded biases. To generalize Aigner and Cain's statement (1977: 177), "To interpret the 'statistical theory of discrimination' as a theory of 'erroneous' or 'mistaken' behavior by [decision-makers such as] employers, as have some economists, is without foundation."

This definition of statistical discrimination has been described as the "strong version of the statistical discrimination hypothesis, typically associated with economists" (Tomaskovic-Devey and Skaggs 1999: 424), as compared to a weaker version, associated with sociologists, allowing for erroneous beliefs and stereotypes to be included within the definition of statistical discrimination. Illustrating this weaker version, a recent sociological review defined statistical discrimination as deriving from "known or assumed differences in competencies between groups" (Stainback, Tomaskovic-Devey and Skaggs 2010: 233). We use the strong version, and concur with the panel of scholars authoring the National Research Council's (NRC) *Measuring Racial Discrimination* (2004) that allowing the definition of statistical discrimination to include potentially biased perceptions and beliefs renders the theory meaningless and empirically indistinguishable from bias. The NRC scholars emphasized this point in the statement below:

> "When beliefs about a group are based on racial stereotypes resulting from explicit prejudice or on some of the more subtle forms of ingroup versus-outgroup perceptual biases, then discrimination on the basis of such beliefs is indistinguishable from the explicit [and non-statistical] prejudice discussed above. Statistical discrimination or profiling, properly defined, refers to situations of discrimination on the basis of beliefs that reflect the actual distributions of characteristics of different groups" (NRC 2004: 61-62).

Even scholars advocating statistical discrimination as an explanation for racial disparities in care accept this proper definition of statistical discrimination. McGuire and colleagues recently defined statistical discrimination occurring only when "providers apply *correct* information about a group to reduce their clinical uncertainty about an individual patient" (2008:2, emphasis added).

Identifying Statistical Discrimination

Although the above definition of statistical discrimination may seem stringent to the point of making it unlikely that this mechanism could ever be positively identified, this is not the case. The key to identifying statistical discrimination lies in scrutinizing its dynamic rather than static predictions. For a host of reasons, the static prediction, that decision-makers base their decisions on the true distributional characteristics of the social categories, is virtually impossible to verify empirically and conclusively. The dynamic predictions of statistical discrimination, that is, predictions about when and how decision-maker behaviors would change under statistical discrimination, not only provide a way to positively identify statistical discrimination, but have documented success at doing so.

Changes in disparate outcomes under statistical discrimination may result of any of three processes. The first process is ***correctional changes***, where the decision-makers' initial erroneous (i.e. bias-based) decisions may be in the process of being corrected via competitive forces and coming into alignment with what is predicted by statistical discrimination. The second process is ***population changes***, where the means or the variances in the hard-to-observe characteristic may have changed for a group in the population. The third process is ***information changes***, where there is a change in the availability of the hard-to-observe characteristics of the target population. For the first two processes, the changes in disparities move towards the disparities entailed by the true distributional differences among groups. In the final process, the level of disparity is related to the availability information concerning the hard-to-observe characteristic. We consider whether and how each of these change processes apply to our empirical context.

Previous scholarship (described in more detail below) has successfully revealed positive evidence for statistical discrimination using the information change process. This process is based on changes in the hard-to-observe characteristic. Consider the implications of having more direct information about the hard-to-observe but valued characteristic. If a decision-maker were to have accurate information about the hard-to-observe characteristic for a particular set of individuals, then net of that hard-to-observe characteristic, group membership should have no association with treatment decisions. Even in the presence of noise in the signal of the hard-to-observe characteristic, if a decision-maker were to interact with a set of individuals who provided signals of their hard-to-observe characteristics with uniform noise (i.e., signal variance was uncorrelated with group membership), then again group membership should have no net effect on that decision-maker's decisions. As decision-makers have more direct information about the hard-to-observe characteristic, and as the signal about that characteristic is presented in a manner uncorrelated with social category, then statistical discrimination predicts lower disparate outcomes net of the hard-to-observe characteristic. This relationship between information and disparate outcomes has been the key to empirically test for positive evidence of statistical discrimination.

For statistical discrimination in labor market outcomes, the hard-to-observe characteristic is usually considered to be some form of worker productivity (Correll and Benard 2006). If an employer were to have more direct and less noisy (or more specifically, noise that is uncorrelated with social category) information about productivity, then net of that information, disparities in outcomes by social group should diminish. The amount of relevant productivity information an employer has about an employee or potential employee is lowest pre-hire, and increases with employee tenure with the employer. Thus, if an employer only engages in statistical discrimination, the association between race and wages, for example, should diminish with employee tenure. Starting with this insight, Altonji and Pierret (2001) tested for such a diminishment, but found the opposite – an *increase* in the association between race and wages with employee tenure. They did find this diminishment in the effect of years of education, suggesting that while statistical discrimination may explain unequal wages by years of education, it is unlikely to explain unequal wages by race. This example of positive evidence for the operation of statistical discrimination in the labor market is for statistical discrimination by educational status and not by racial category. The racial disparities in wages are likely the result of other non-statistical mechanisms.

Clearly, statistical discrimination can be positively identified with the appropriate research design. Currently, there are few examples in the literature of such designs. One of the obstacles to performing empirical research using an appropriate research design to rule-out statistical discrimination is the need for longitudinal data of disparate outcomes by the same decision-makers. Attributing disparate outcomes to decision-maker behaviors, rather than other mechanisms, is empirically difficult. Doing so over time for the same decision-makers is all the more difficult. Thus, the dearth of positive evidence for statistical discrimination is less an indictment of the theory and more a testament to the difficulty of disentangling that mechanism from others also contributing to disparities.

To falsify statistical discrimination in explaining racial disparities in care, this paper leverages the predictions of statistical discrimination regarding changes in disparate outcomes under the three change process described above. We test for changes in discriminatory behaviors in an empirical setting where statistical discrimination change processes would predict only reductions in disparities. In this setting, any measurable *increases* in discriminatory behaviors *cannot* be attributable to statistical discrimination. In terms of our earlier metaphor, we test for changes in new babies in a setting without storks. The disparity we investigate is racial disparities in patient care, and the empirical setting is the first two years of medical school training.

Racial Disparities in Patient Care

U.S. health disparities by race, where white Americans experience significantly better health outcomes than black Americans, are pervasive and enduring, and have a myriad of complex causes (Agency for Healthcare Research and Quality 2008). One troubling contributor is that U.S. physicians

treat patients differently by race (Institute of Medicine 2003). The existence of racial disparities in patient care delivery by physicians has been well-documented and established (Institute of Medicine 2003; van Ryn 2002). Notably, these disparities are independent of the race of the physician. That is, both black and white physicians generate similar racial disparities in care (Chen et al. 2001).

Unfortunately, these disparities have been disturbingly resistant to change despite more than a decade of awareness and many efforts aimed at addressing the issue (Devi 2008; Gross et al. 2008; Orsi et al. 2010; Pletcher et al. 2008; Vaccarino et al. 2005). The causes of care disparities remain elusive (Klonoff 2009), but many scholars have posited statistical discrimination as an important, and possibly the primary, explanatory mechanism (e.g. Chandra and Staiger 2010; Chin and Humikowski 2002; Lutfey and Ketcham 2005; McGuire et al. 2008; Balsa and McGuire 2003). These studies reveal associations consistent with statistical discrimination, but also consistent with other mechanisms. They have neither provided unambiguous evidence for statistical discrimination, nor demonstrated the absence of other discriminatory mechanisms.[1] We describe below the design of our study, aimed at being able to falsify the statistical discrimination explanation for racial disparities in care.

RESEARCH DESIGN

This study uses a longitudinal quasi-audit of medical students during the first two years of medical school to show that statistical discrimination is unlikely to be either the only or primary explanatory mechanism for racial disparities in patient care. In this section, we describe the context of this empirical study – the first two years of medical school, the nature of the quasi-audits, and statistical discrimination's predictions of first to second year changes in medical student behavior as revealed by these quasi-audits. Whereas constant or decreasing disparities between the first and second year would be consistent with statistical discrimination, increasing disparities would be inconsistent.

A Strategic Research Site: The First Two Years of Medical School

The decision-makers (potential discriminators) in our study are medical students. The potential targets of discriminatory behavior are standardized patients (SPs) – actors trained to present a scripted clinical case to medical students. The use of standardized patients in medical school is a long-established pedagogical technique (Barrows 1971, 1993) that has grown significantly with the 2004 addition of 10 clinical SP case encounters as a part of the U.S. Medical Licensing Exam. Race-varying SPs presenting a clinical case that does not involving any race-relevant pathology create a natural audit-study, allowing a good measure of differential treatment (NRC 2004; Quillian 2006:303). Although physicians, not medical students, are the decision-makers contributing to actual care disparities, studies using SP case-encounters have documented racial disparities in medical student outcomes (Colliver et al. 2001; Beach et al. 2007).

[1] In a notable exception, Chandra and Staiger (2010) do falsify the Beckerian "taste" bias mechanism as an explanation for racial disparities in the treatment of Medicaid patients having experienced heart-attacks.

Medical school training in the U.S. follows a highly institutionalized four-year structure (Cooke et al. 2006). Whereas the first two years of medical school are characterized by strong cohort unity and traditional classroom-based pedagogy, the last two years are independent and apprenticeship-oriented. The first two years of the medical school curriculum focus on classroom-based and laboratory learning, with a cohort of medical students taking almost all the same classes in the same order. During these first two years, medical students have limited direct clinical encounters with actual patients. Students in the final two years follow individualized schedules and have individualized patient care experiences. Medical students' limited direct clinical experiences during their first two years undermine claims that increases in disparities may be attributable to statistical discrimination.

In addition to the limited exposure to clinical experiences, the formal curriculum of the first two years of medical school is also relevant. The medical profession has responded to the finding of physician-generated disparities by altering medical school curricula. In 2002 the Liaison Committee on Medical Education (LCME 2008) added the requirement that all member medical schools include cultural competence skills training (ED-21), and that all member medical schools provide instruction on the existence of racial disparities in diagnosis and treatment (ED-22). One purpose of these requirements is to reduce physician-generated disparities (Betancourt 2006; National Partnership for Action 2010). The specific structure and format by which medical schools meet these requirements are left to the discretion of each individual medical school. To keep their accreditation, medical schools have worked to ensure that the explicit lessons provided during medical training do *not* lead to racial disparities in care.

Longitudinal Natural Experiment: Repeated Quasi-Audits

This study scrutinizes the changes in care disparities by three cohorts of medical students between their first and second years of medical school. The uniformity, short time-span, lack of subject attrition, and limited clinical exposures of the first two-years of medical school; paired with the common practice of performing natural quasi-audit studies on the students makes this setting a "strategic research site" (Merton 1987) for investigating changes in discriminatory behavior.

When black and white SPs are assigned randomly to a cohort of medical students engaging in SP case-encounters, disparities may be measured in differences in the encounter outcomes between black and white SPs. The random assignment creates the natural experiment where medical student characteristics are unlikely to be associated with whether they interact with a black or white SP. The black and white SPs portraying the same clinical case defines the quasi-audit study where race is an exogenous manipulation of otherwise identical stimuli presented to decision-makers. We track the disparities revealed by this quasi-audit from the case-encounters performed by all first and second year medical students to test for changes in racial disparities in care in a longitudinal natural experiment.

The core distinguishing feature of audit studies versus other methods of measuring disparities is the use of paired testing (Fix and Turner 1998: 11). Paired testing allows an objective answer the question whether one (or more) manipulated dimensions (such as race [Yinger 1986], gender [Neumark et al. 1996], criminal history [Pager 2003]), rather than any other characteristic or trait, gives rise to disparate treatment by a decision maker. These specific dimensions are scrutinized in isolation by exposing decision-makers with paired versions of the kind of stimuli they experience during the normal course of their decision-making process. These paired stimuli, often actors, but sometimes simpler stimuli such as resumes, are trained or designed to be observationally equivalent except along the manipulated dimensions. A great advantage of audit studies is they allow a measurement of discrimination by decision-makers when they are making the actual decisions that result in the disparate outcomes being studied. Examples include decisions to invite job applicants to be interviewed or hired to actual jobs for studying job segregation (Bertrand and Mullainathan 2004; Pager 2003); or decisions to show, rent or sell actual available real estate for studying housing segregation (Yinger 1986). Audit studies are currently one of the best ways to measure actual racial discrimination (NRC 2004; Quillian 2006). Even critics of the auditing method acknowledge that audits are "the only objective means of detecting discriminatory treatment" (Siegelman via Fix and Turner 1998: 3).

Our study uses paired testing with actors, but differs from traditional audit studies in several important ways. First, the behavior of medical students during standardized patient encounters do not contribute to actual disparities in health. Relatedly, the medical students – not yet being actual doctors – know they are interacting with actors and not actual patients, and that their performances in these encounters are being graded. These differences behoove caution in drawing a direct line between our findings and the mechanisms underlying disparities in actual patient care.[2]

Some of the differences between the nature of our audit study and traditional audit studies make our study a better and more conservative test of disparities. One of the strongest critiques leveled against audit studies involving actors is that the actors are aware of the nature of the study and may subtly or unintentionally engage in behaviors that make the finding of differences more likely (Heckman and Siegelman 1993; Quillian 2006). However, our study is a double-blind audit: neither the medical students nor the actors are aware that the data from these encounters are used to investigate racial differences in care. Thus, it is highly unlikely that the actors work to confirm differences. In addition, the fact that students know that their performance in these encounters affects their grades introduces a level of accountability not commonly present in traditional audit studies. When decision-makers are aware that

[2] The USMLE added standardized patient encounters to the exams required to earn a medical degree in part because these encounters increase the fidelity of assessments of medical students' likely performance as a physician beyond the previous set of exams.

their decisions are being externally scrutinized, this accountability may reduce the biases manifest in their decisions (Russo et al. 2000; Tetlock and Mitchell 2009). If this tendency holds in our study, then our design would make it harder to detect disparities, and thus act as a conservative test for disparities.

Audit studies have known limitations in addition to self-fulfilling behaviors by actors (Heckman and Siegelman 1993; NRC 2004:108-114; Quillian 2006:304). Most of these limitations do not apply to the current study. The accuracy concern about audit studies (often for employment or housing settings) is that the same target of study does not receive multiple audits from both (or all) conditions. Our study uses many audits by both black and white SPs of the same cohort-year of medical students to measure disparities at that level. Similarly, our study is largely immune to the concern that audit studies' measures of bias are localized to a particular event (e.g., a job interview), which may represent only a small part of the phenomenon being studied (e.g., employment discrimination). We are explicitly focused on studying racially-biased outcomes from clinical encounters, and not other aspects of racial health disparities. Indeed, the fact that these encounters are a required part of students' formal medical training greatly enhances the ecological validity of our study relative to explicitly lab-based studies of discrimination (cf. Tetlock and Mitchell 2008:14). One concern about audit studies is that measures of bias at a particular site (e.g., a firm conducting a job search or looking for a renter) may not generalize to the market or region. Our study is essentially a quantitative case study, and shares the generalizability limitations of case studies. That is, although we are able to discern with exquisite detail the changes in disparities revealed in our setting, we cannot make definitive claims that such dynamics may be expected to occur in all such settings. That said, we also have no reason to believe the medical school under investigation to be an atypical medical school in terms of how it affects the discriminatory behaviors of its students.

Even the significant concern of auditor heterogeneity is only a minor concern of our study. Despite being trained to behave uniformly, the actors cannot behave exactly alike. In this study, the actors are standardized patients (SPs) – trained not only to conduct the audit, but to evaluate the performance of the auditee, the medical student. If SP heterogeneity related to evaluations is also correlated with SP race, then measured racial disparities in care could actually be the result of this correlated heterogeneity. This explanation was the one given by Colliver and colleagues for their empirical finding of consistent and significant disparities (2001: 12) in a cross-sectional studies of fourth year medical students. Although the heterogeneity could contribute to findings of bias, our main concern is identifying *changes* in bias between the first and second years of medical school. For auditor heterogeneity to contribute to any identified *trends*, the heterogeneity effect would have to be different not as a function of the race and experience or tenure of the auditor, but as a function of tenure of the medical student the auditor is evaluating. So the concern would not be that black SPs might evaluate medical students more harshly than white SPs (which would have an effect on a difference, but no effect on a trend), but that black SPs'

harshness of evaluations might increase (or decrease) for second year medical students relative to the evaluations given by white SPs.[3]

Statistical Discrimination and Changes During the First Two Years of Medical School

Above, we described three types of change processes consistent with statistical discrimination (correctional, population and information). Here, we take each change process in turn and consider their implications for our research setting.

Correctional Changes

Correctional changes in disparities occur when perceptual errors and biases are competed away in the market. The result of correctional changes under statistical discrimination is changes towards treatment disparities entailed by the actual distributional differences among groups within the population. Although the pre-correction level of disparities may be more or less extreme than the statistically defined post-correction levels, the end point of such corrections is exactly the single statistically defined level.

For these correctional changes to occur, decision-makers must be participants in a competitive market. Even advocates of statistical discrimination explanations of racial disparities in care acknowledge that healthcare is not a good example of a competitive market (Balsa and McGuire 2003: 95-96). For any competition in healthcare to change physician care via correctional changes in statistical discrimination, physicians must experience some costs when treating patients in a manner that is not statistically justified. These costs could derive from the misdirection of scarce or expensive resources, negative patient outcomes, reputational costs, loss of patients to other (more statistically appropriate) physicians, or other costs from treatment behaviors deviating from statistical optima. These costs affect both the variance-based and means-based forms of statistical discrimination in similar ways and with similar implications for the purposes of this study.

If such changes occur at all, correctional changes require interaction over time with other informed actors in the competitive market. Even if such correctional changes can occur, they are unlikely to explain changes in the behavior of medical students between their first and second years of medical school. These students cannot be considered market participants. Their limited observations of clinical encounters rarely involve any repeated encounters with the same patients. As a result, first and second

[3] Although racial effects on changes in evaluation harshness may seem unlikely, something of a similar nature has been found. Simons and colleagues (2007) found that black employees rated behavioral integrity violations by their managers more harshly than did their white counterparts. If some similar kind of violation (e.g., increasing emotional detachment by the medical student [Mizrahi 1996; Spiro 1992]) is more common among second year students than first year students, then racial differences in responses to those behaviors could appear as a trend in disparities. We address the concerns of auditor heterogeneity and racial differences responses to first versus second year students directly in our analysis.

year medical students do not directly observe or experience the kind of feedback required by the correctional change processes. Changes in disparities exhibited by medical students between their first and second years are unlikely to be explained by correctional changes under statistical discrimination.

Population Changes

Racial associations with the means and variances of hard-to-observe characteristics may exist and may also change over time. Any such changes would likely be very gradual. The chance that population change processes explain changes in disparities from statistical discrimination between the first and second years of medical school is exceedingly low. This chance approaches zero if the observed changes are consistent across cohorts in a non-contemporaneous multi-cohort study.

Information Changes

The process of information changes has slightly different implications for means-based and variance-based statistical discrimination in our setting. For means-based statistical discrimination, information changes should yield changes in disparities. An important purpose of medical education is to train students to be effective care-givers. It is reasonable to hope that medical schools improve medical students' abilities to detect and identify hard-to-observe patient characteristics that are diagnostically relevant to the patient's health. If such changes do take place during medical training, then medical students should have greater access to hard-to-observe characteristics with greater training. This is identical to having more hard-to-observe information. As a result, net of those hard-to-observe characteristics, characteristics like race should have *less* of an association with care outcomes. In our specific setting of performance in standardized patient encounters, these hard-to-observe characteristics relevant to health and diagnosis are held constant across race-varying standardized patients within each clinical case, and thus the effects of these characteristics are already accounted for by design. Through the information changes process, means-based statistical discrimination predicts that disparities should be *reduced* between the first and second year of medical training.

For variance-based statistical discrimination, the information process may or may not yield changes in disparities in our setting. In variance-based statistical discrimination, disparities come from race-specific differences in the variance of the health signals generated by patients. Given our use of standardized patient encounters to measure disparities, it is unlikely disparities from variance-based statistical discrimination would be present at all. The SP training process ensures the SPs provide the medical students the same clinical information in the same manner regardless of SP race. The fact that the medical students are aware that they are interacting with actors trained to present scripted symptoms and responses to physician questions further reduces any possible expectations of race-associated noise in these signals. (It is worth reiterating that any physician differences in expectations about or interpretations of patient signals is a perceptual error that is inconsistent with the definition of statistical discrimination.

Statistical discrimination is about the correct and true properties of the groups themselves, not how those properties may be differently perceived or interpreted.) Therefore, based on the characteristics of our research setting, variance-based statistical discrimination is unlikely to be present in either the first or second year, and thus, unlikely to change. If, for some other reason, there were variance-based statistical discrimination in this setting, disparities would be likely to decrease between the first and second year of medical training for reasons similar to the decreases predicted by the means-based variant of statistical discrimination. Better trained students should become better at eliciting health signals from their patients, and not be as subject to the "natural" variances of different groups in generating health signals. So the information change process would predict either no change or a reduction in disparities between the first and second year of medical school.

Implications

Considering the three processes by which statistical discrimination (in both its means-based and variance-based forms) would predict changes in disparities (correctional, population, and information), there should be either no change or a *decrease* in the disparities measured via SP encounters from the first to the second years of medical school. As a corollary, *if we were to observe any significant increase in disparities between the first and second year of medical school, this increase cannot be due to statistical discrimination.*

METHODS AND DATA

Empirical Setting

The Orchard School of Medicine (OSM, a pseudonym) curriculum has each individual medical student complete two similarly-structured standardized patient case-encounters during their first two years of medical school. The first-year case involves students taking the medical history (Hx) of the SPs, and the second-year case involves both a medical history and physical exam (HxPE).[4] Both the first and second year cases were designed such that there is no medical reason for differential treatment based on patient's race. The class size at OSM is usually a little more than 100 students.

First and second year medical students at OSM observe physicians performing outpatient care for several hours once every two weeks. First-hand student experiences with actual clinical encounters and outcomes could conceivably affect student behaviors in a manner consistent with correctional changes under statistical discrimination. As discussed above, correctional changes require participation in the

[4] The first year History (Hx) case and the second year History and Physical Exam (HxPE) case are comparable for the history-taking component present in both cases. To ensure valid comparisons in our analysis, we use only those outcomes from the cases for which the same sets of behaviors are evaluated in both settings: the history itself, and the patient-physician interaction behaviors. We also include the overall subjective rating of patient satisfaction, as it has been seen as an important part of racial disparities in care (see Institute of Medicine 2003:574-575; van Ryn 2002:I-146). OSM provided these outcome measures. As discussed below, we also coded one cohort's encounter videos to address validity concerns.

competitive market over time to observe and experience the costs and benefits of statistically inappropriate and appropriate care, respectively. The low frequency and duration of students' clinical exposures during their first two years limits their exposure to individual patient follow-ups and their ability to detect the kinds of benefits and costs entailed by more or less statistically appropriate care. Absent direct experience with the (questionably [see Balsa and McGuire 2003: 95-96]) competitive health care market forces, there is no statistical basis for correctional changes in clinical encounter behaviors by patient race.

The data in this study were collected after cultural competency training became institutionalized as a requirement of medical education. At OSM, cultural competency training takes place as a unit (in the form of several hours of lecture time) within one of the required first semester courses of the first year. After the successful completion of that unit, there is no requirement for formalized follow-up or reinforcement of cultural competency training during students' remaining time at OSM.

The first year OSM students' first SP case-encounters take place towards the end of the spring semester – well after they have all completed their cultural competency training. This temporal structure is a benefit to our research design. All SP case-encounter observations use students who have completed the same school-required cultural competency training. Any effects of that training should be present across all observations.

Data Sample

This study uses data collected during the first and second year SP case-encounters within the regular curriculum at OSM. Since 2006, OSM has kept records of the specific student-SP pairings – a requirement for our analyses. Our data come from three cohorts of students in OSM's M.D. program (the classes of 2009-2011) with both first and second year encounters taking place between 2006 and 2009.

For both the first and second year case-encounters, many SPs present the identical case to the entire cohort. For this reason, OSM employs a variety of actors of varying race/ethnicities to present each case. Actor schedules, and no characteristics related to the medical students, determine whether a particular student sees a black or white SP. The structural independence between student characteristics and SP race provides the serendipitous randomization underlying this natural experiment. We compare the student behaviors when interacting with black SPs to those when interacting with white SPs. We exclude cases where there was no SP race data, where the SP was neither black nor white,[5] or where the encounter was a repeat of one already performed by the student (as is sometimes requested either by the

[5] The actors hired as standardized patients self-identified the racial categories they can portray in their acting roles. Very few standardized patients identified as being neither white nor black. We repeated our analyses with these excluded SPs grouped with the black SPs (white/non-white) and with these excluded SPs grouped with the white SPs (black/non-black). We found no differences in direction or statistical significance of our results in these variations.

student or the medical school). These constraints yield 582 SP case-encounters for our analysis. Table 1 details the number of observations obtained from each cohort and case used in this analysis. Forty-six actors (38 white and 8 black) presented the first and second year cases to the three medical student cohorts studied.

[TABLE 1 HERE]

Based on the data from these SP encounters, we present two sets of analyses. The first set of analyses are performed on all three cohorts using the data provided by OSM. The findings from these initial analyses are confirmed in a second set of analyses are performed on one cohort (the class of 2011) based on the results of independent coders who coded video recordings of the encounters. The latter set of analyses address some of the design and data limitations otherwise present in the study.

3-Cohort Analysis Dependent Variables: Standardized Patient Encounter Outcomes

Following each standardized patient case-encounter, the medical student's performance is evaluated by the SP against a checklist of objective behaviors and actions. In addition, SPs are asked to rate subjectively the medical student's performance in terms of their satisfaction as a patient. We use three outcome measures common to the two cases used for first and second year medical students. These outcome measures are:

History: Did the medical student ask all the questions necessary to assess the patient's complete medical history? This measure is the percent of questions asked from a checklist of approximately 60 questions. Examples include that for every symptom the patient names, the medical student is supposed to ask about that symptom's severity, the time-of-day when it tends to occur, and the impact the symptom has had upon the patient's life, among others. An example of this checklist is given in the appendix.

Patient-Physician Interaction (PPI): Did the medical student enact the 14 behaviors emphasized in student training and shown to support a successful clinical encounter (e.g., introducing herself by name, calling the patient by name, maintaining eye-contact)? This measure is the percent of behaviors noticed by the SP from a checklist of established behaviors.[6] The checklist appears in the appendix.[7]

Patient Satisfaction: A two-item subjective evaluation by the SP of whether she or he would return to the medical student for care, and whether he or she would recommend the medical student to a friend or family member seeking care. Both questions use a 5-item Likert-type scale. These are coded 0-4, summed, and divided by 8 for an outcome that ranges from zero to one in one-eighth increments.

[6] The patient-physician encounter is a highly institutionalized component of the care-giving process (Heritage and Maynard 2006:363), and is studied extensively for behaviors associated with improved medical outcomes (a brief review in Heritage and Maynard 2006:365; specific examples in Smith 2003).

[7] The response options on the PPI checklist for the second year case for the 2011 cohort was altered from binary yes/no to a 4-item (1-4) Likert-type option. The same 14 items in the appendix appeared on all checklists. We test for effects from this scoring change using the methods discussed below, including adding a new dummy variable *PPICHANGE* (1=second year case for the 2011 cohort, 0 otherwise).

Video Coding Analysis Dependent Variables: Non-verbal Behaviors and Demeanor

The outcomes provided by OSM for the two types of cases were the students' item scores from their entire encounters. In the first year, the history-only encounters were typically 20 minutes long, while in the second year, the history plus physical exam encounters were about an hour long. Even though we use comparable outcome measures for the two cases, the differences in the cases themselves and the amount of time the students had to demonstrate the evaluated behaviors are problematic confounders for our analysis. To address these and other design and data concerns of the study, we also analyze results from an independent coding of video recordings of the class of 2011's SP encounters from their first and second years.

Twelve coders, naïve to the study's research question, were trained to code only the history-taking portion of video recordings of the SP case encounters for both the first and second year case encounters – both approximately 20 minutes long. Each video was coded by an average of 3 coders.[8] Coders were scheduled so they would not code videos from the same medical student more than once.

Because we could not use the same checklists as the above analysis (which are based on the complete encounter), coding focused on non-verbal behaviors and demeanor shown by previous scholarship to be associated with expressions of racial bias and/or empathy in social interactions. These items include *smiling* and *leaning toward the SP* (McConnell and Leibold 2001:440); and a set of positive adjectives describing the medical student's *apparent demeanor*: likeable, warm, friendly, and pleasant (Richeson and Shelton 2005).[9] Each item was scored using 7-item Likert-type response options. The exact coding instrument wording is provided in the appendix.

Confirmatory factor analysis showed the two non-verbal behaviors to be unique, and the four demeanor adjectives loading onto a single factor, allowing them to be averaged into a single positive demeanor index (Cronbach's alpha = 0.92). Inter-rater reliability (ICC(2,k) [Shrout and Fleiss 1979], as we use the average of the coders' ratings) for these three items were all large and strongly significant (smiling: 0.75, leaning: 0.73, positive demeanor: 0.75). We reverse-coded each of these items, so their 1 to 7 range corresponded to never occurred to always occurred, respectively.

Independent Variables & Controls: Case, Actor & Student Cohort Characteristics

Case characteristics. The first and second year encounters use different clinical cases. The first year has a shorter encounter where the standardized patient's chief complaint is that of abdominal pain, and the medical students take a complete history (Hx) of the patient, but do not perform a physical

[8] Coders self-identified their racial categories. Five self-identified as Asian, 3 as white, 2 as Hispanic or Latino/a, 1 as black, and 1 as other. We scheduled coders so no video was coded solely by white coders.
[9] The original instrument included other behaviors and adjectives which were excluded because the medical students either always (e.g., eye-contact) or never (e.g., crossed arms) exhibited them. Details about the full set of items from this instrument are available by request.

examination. The second year has a longer encounter where the standardized patient's chief complaint is that of a chronic cough, and the medical students conduct a complete history and physical examination (HxPE). We use a dummy variable for year of medical school (*YEAR:* 0 for the first year, and 1 for the second year) in our analysis. This dummy variable allows us to test for year-specific associations with the outcome variables, whether such associations arise from observable or unobservable year characteristics. Because the year variable is identical to a dummy variable for the case of medical SP encounter (the Hx or the HxPE cases), it also controls for any case-specific effects on the means of the outcomes. The *YEAR* variable serves as the basis for our interaction term – the focus of our analysis.

Actor characteristics. The race of the standardized patient is central to our analysis. As we have restricted our analysis to black and white standardized patients, a single binary variable (*SPWHITE:* 0 for black SPs and 1 for white SPs) codes standardized patient race. In addition to actor race, we have data on actor sex (*SPFEMALE*), age (*SPAGE*), and experience as an SP (*SPEXPERIENCE*) as measured by the count of encounters they had performed at the time of the encounter. This latter set of actor characteristic variables serve as controls for actor effects in some of our estimation models. In other estimation models, we perform a fixed-effects analysis that essentially creates a dummy variable for the standardized patients to control for all standardized patient characteristics, whether observed or unobserved.[10]

The key variable of interest is the interaction between year of medical school training (*YEAR*) and standardized patient race (*SPWHITE*). This interaction term (*SPWHITE X YEAR*) measures the degree to which the effect of standardized patient race changes between the first and second year of medical school. Given our definition of the case and *SPWHITE* variables, a significant and positive coefficient for our interaction term would indicate a significant increase (decrease) in the outcomes of second year students when interacting with white (black) patients relative to the first year outcomes. Similarly, a significant and negative coefficient would indicate a significant decrease (increase) in the outcomes of second year students when interacting with white (black) patients. Such an effect could indicate the growth or diminishment of disparities in outcomes between the first and second year of medical school.

Student cohort characteristics. A dummy variable (*COHORT*) for the three medical school cohorts controls for any observable or unobservable differences in the cohorts such as effects of the composition of the cohort in terms of race, bias, or other characteristics.

Medical student characteristics. Student information was extremely limited in the extant data.[11] Only student sex was provided. Although the random assignment of student to SP should obviate spurious

[10] In the fixed-effect models, the SP race dummy, *SPWHITE*, is necessarily omitted. This omission does not affect the identification of the key interaction term.

[11] From school-wide demographic data published by OSM, we know OSM had a significantly larger composition of black medical students (10%-12%) than contemporaneous national averages (7%), but a similar composition of white students (60%-64%) as contemporaneous national averages (61%-63%).

findings from student heterogeneity, we also test this possibility directly in our analysis based on the coded interaction videos. We used the class of 2011 cohort videos to code medical student race to allow an analysis of concordance (cf. Cooper-Patrick et al. 1999).[12] As we have only black and white SPs in our sample, medical students were coded as black (14 students), white (66 students), or other (21 students) for the purpose of analyzing the effects of concordance. The available student demographic information is summarized by cohort in the lower panel of Table 1.

Student-patient concordance. Because sex and race may interact in the production of disparities in patient care (e.g., Schulman et al. 1999), we model concordance in a manner to account for this potential interaction. There are four possible gender pairings of medical students and SPs. We model this using three binary indicator variables: *SPFEMALE*, *STUDENTFEMALE*, and *BOTHFEMALE*. For racial concordance, because some medical students are coded as neither black nor white, we include the following five binary indicator variables: *SPWHITE*, *STUDENTWHITE*, *STUDENTBLACK*, *BOTHWHITE*, and *BOTHBLACK*. Finally, to allow for interactions in the sex and race concordance effects, we add a *MATCHSEXANDRACE* variable that is one when the student and SP match on *both* sex and race dimensions, and zero otherwise. We also include SP age and experience as before. Because we can analyze concordance effects in our class of 2011 cohort for any of our dependent variables, we also perform a supplemental analysis of the history, PPI, and patient satisfaction outcomes with this same model as a robustness check.

Estimation Strategy

We estimate changes in racial disparities in care between the first and second year of medical school using linear regression and linear regression with fixed-effects. The main advantage of the regression approach to estimating trends in disparities is the ability to control for idiosyncratic effects from different SPs, different cases and different medical school cohorts, gender concordance between the SP and the medical student, and for the one cohort with coded videos, racial concordance as well. Estimating effects within-cohort obviates any cohort-specific effects, and randomization to exposure takes care of most of the concerns regarding individual differences among medical school students. The fixed effects analysis is the most conservative way to control for SP heterogeneity and idiosyncratic ratings.

The conservative nature of the fixed effects analysis means that estimates of some actual effects are potentially attenuated by the many dummy variables used to represent the 46 SPs. As the number of groups in a fixed effects analysis grows, estimates may become inconsistent (Nickell 1981). We also perform a simpler regression using the actor-level controls described above.

A stylized version of the general regression model we use is as follows:

Similar school-wide statistics show OSM as having compositions of women (48%-50%) comparable to contemporaneous national averages (48%-49%).

[12] To keep the 12 video coders naïve to our research question, other video coders performed the sex and race coding of the medical students from the videos.

OUTCOME = *YEAR* + *SPWHITE X YEAR* + [*COHORT*] + [SP controls] + [Student controls] + [Concordances]. The *COHORT* dummies are included for analyses of the three cohorts, but not for the single cohort. The SP controls include *SPEXPERIENCE* (present in all models), *SPWHITE*, *SPAGE*, and *SPFEMALE*, or SP fixed effects. Student controls include *STUDENTFEMALE* (in all models), and for the class of 2011 cohort, *STUDENTBLACK* and *STUDENTWHITE*. Concordances include *BOTHFEMALE* (in all models), and for the class of 2011 cohort, *BOTHBLACK*, *BOTHWHITE*, and *MATCHSEXANDRACE*. When the Outcome is the Patient-Physician Interaction score, we include a *PPICHANGE* dummy variable for the second year encounter of the class of 2011 to control for the changing in the scoring for that encounter.

Again, the key variable of interest for our study is the interaction term, *WHITESP X YEAR*. A significant positive (negative) coefficient for that variable indicates that the effect of the race of the SP on the outcome variable increases (decreases) between the first and second year; that is, a significant increase (decrease) in racial disparities in care as measured by *OUTCOME*. A significant positive coefficient on this term would reveal a significant increase in disparities inconsistent with statistical discrimination.

RESULTS

Table 2 provides the counts, mean scores, and standard deviations for all the six outcomes (three graded encounter outcomes and three video coding outcomes) by medical school year and the race of the SP. The "Differences" column in Table 2 provides the outcome means for white SPs minus those for the black SPs for each year, and the standard error for each difference. The differences in these differences, divided by their pooled standard errors, provides a t-statistic for whether these differences are significantly different from each other. Five of six outcomes (all but history-taking) show significant *increases* in disparities between the first and second year of medical school. These increases challenge the statistical discrimination explanation of racial disparities in care. In addition to these means-based differences, our regression analysis tests more rigorously whether disparities in medical student behavior by the race of the patient increases between the first and second year of medical school.

[TABLE 2 HERE]

Table 3 presents estimated regression coefficients for both the SP controls and fixed effects models with the available student controls and concordance variables. The coefficient of the interaction term, *WHITESP X YEAR*, estimates trends in disparities by year. We find no significant trend in disparities for History, but significant increasing trends in disparities for both Patient Satisfaction (marginally significant in the fixed effects model and strongly significant in the simpler model) and PPI (strongly significant in both models). Both results again show the effect of SP race is significantly larger for second-year medical students than for first-year medical students in the direction of *increasing* disparities.

[TABLE 3 HERE]

Using these estimates, we calculated the predicted values for each of the three outcomes from first and second year medical students encountering white and black standardized patients. The results are plotted in the three panels in the left column of Figure 1 – one panel for each of the three outcomes. The lighter gray lines reveal the first to second year trends in outcomes from encounters with white SPs, and the darker lines, trends in outcomes from encounters with black SPs. The top panel of Figure 1 illustrates these predicted trends for the history outcome. Although there is a general trend towards higher scores from the first year to the second, the lack of any difference in outcomes by SP race is apparent. The lower two panels look very much like each other but very different from the top panel. The lower two panels show no outcome differences by SP race in the first year, but large differences appear in the second year. These differences come from a significant decline in outcomes for second year medical students interacting with black SPs.

[FIGURE 1 HERE]

Examining the non-verbal outcomes derived from independently-coded videos of the SP encounters for the class of 2011 cohort, we find evidence of the same increase in disparate behaviors. This analysis provides three main improvements to supplement the 3-cohort analysis above. First, this analysis controls for racial concordance effects between the SPs and the medical students, as well as the intersection of sex and race concordance. Second, this analysis uses outcomes based on the same initial 20-minute history taking procedure that was common to both first and second year encounters (rather than scores based on behavior over 20 minutes in the first year and over an hour in the second). Third, the outcome scores in this analysis are based on the verifiably consistent ratings of a racially diverse team of independent judges, rather than the individual scores provided by the SP. Table 4 reports the estimated coefficients for these outcomes. Although we do not find significant effects for all three outcomes, the *WHITESP X YEAR* measure of disparity increase is significant for medical students' positive demeanor.

[TABLE 4 HERE]

The three panels in the right column of Figure 1 plot these models' predicted outcomes . For all three outcomes, the slope of the change in medical student behavior when interacting with white SPs is positive between their first and second year. The slopes showing the change in medical student behavior when interacting with black SPs are always less positive, and in two cases, negative. The pattern is consistent across outcomes, and the differences between the two slopes reaches significance for one of the three outcomes: medical students' apparent positive demeanor. The same pattern of increasing disparities in the graded encounter outcomes is evident in the demeanor of medical students as coded by a team of racially diverse judges based only on the initial 20-minute history-taking component of the SP encounters.

Because this class of 2011 analysis allows a more detailed examination of concordance effects than was available from our 3-cohort analysis, we test whether our findings from the 3-cohort analysis are

attributable to racial concordance by repeating the analysis for the three graded encounter outcomes provided by OSM. The three rightmost columns of Table 4 presents our estimates. As before, the key variable, *WHITESP X YEAR*, is positive and strongly significant for the PPI outcome, marginally significant for the Patient Satisfaction outcome, and not significant for the history outcome. Consistent with previous research (e.g., Cooper et al. 2003; Cooper-Patrick et al. 1999; LaViest and Nuru-Jeter 2002), black patients do report higher satisfaction when interacting with black physicians. Also consistent with previous scholarship (e.g., Schnittker and Liang 2006), these concordance effects do not explain the disparate outcomes. Despite the presence of concordance effects, the increasing disparities we find are neither attributable to nor diminished by concordance in sex, race, or the interaction between the two.

DISCUSSION AND CONCLUSION

Consistently and robustly, we find a measureable increase in the disparities exhibited by medical students from their first to second years. The three change processes consistent with statistical discrimination predict either no changes or reductions in disparities between the first and second year of medical school. Therefore, the observed growth in disparities likely derives from sources other than statistical discrimination. We do not claim that statistical discrimination is wholly absent among the mechanisms underlying racial disparities in care. We simply point out that based on our evidence, it is unlikely to be the only or even the primary form of discrimination generating these disparities.

This study has a unique research design. There have been longitudinal studies of changes in discriminatory outcomes, but no longitudinal studies that provide the experimental clarity of an audit study. There are audit studies giving clean estimates of discriminatory outcomes, even in medical settings using standardized patients, but none are panel studies to allow an investigation of changes within cohorts. Our study combines the strengths of the audit and panel study designs in a natural experiment to provide singular scrutiny on the development of disparities in service delivery within a profession.

Early discrimination research often attributed disparate outcomes to bias based on findings merely consistent with bias, though lacking positive evidence for a bias mechanism. Inferring bias mechanisms from the significance of a race coefficient or a similar residual racial gap after including controls is now a justifiably deprecated practice (NRC 2004: 121-122; Pager and Shepherd 2008:184). Merely consistent findings are also not sufficient to make claims of statistical discrimination.

Statistical discrimination explanations for a host of disparities beyond the labor market proliferate despite the lack of any positive evidence for statistical discrimination in those realms. The bar need not be higher for statistical discrimination as compared to bias-based discrimination, but certainly it should be no lower. Positive evidence for statistical discrimination requires more than mere plausibility and either evidence of distributional differences in the population (Pager and Karafin 2009), or a falsification of Beckerian taste discrimination (e.g., Chandra and Staiger 2010; Siniver 2011). Statistical discrimination

may be a prepossessing theory, in explaining inequalities without requiring individual bias, but this is not a reason to privilege this mechanism over others.

In this environment where there are many candidate mechanisms for the generation of disparities by professionals, eliminating one candidate mechanism from consideration represents useful progress. Our study has ruled-out statistical discrimination from consideration as the primary mechanism generating racial disparities in patient care. The significant changes in disparities we document by first and second year medical students are inconsistent with statistical discrimination explanations. What explanations are consistent with our findings? Our findings are consistent with at least two explanations, although we cannot positively identify either of these two mechanisms as being present.

One explanation is that the disparities are based on cognitive biases present among the medical students. These biases, whether explicit or implicit (e.g., Green et al. 2007; Sabin, Rivara, and Greenwald 2008) affect care and are relatively stable (Cunningham, Preacher and Banaji 2001), and do not change between the first and second year. The observed increase in disparities results from some characteristic of the second year encounter that triggers the manifestation of bias – a characteristic that is absent in the first year encounter.

A second explanation is that the disparities are based on cognitive biases that do change between the first and second year of medical school. Because of the general stability of such biases, this explanation requires that medical students are learning or acquiring these biases between the first and second year of medical school. Although first-hand experiential learning through observing effects of treatment decisions on longer-term patient outcomes and behaviors is rare and unlikely for these medical students, learning through observation of actual physician behaviors is not. Students can and likely do learn clinical encounter behaviors by observing the behaviors of physicians during these clinical exposures. Adopting behaviors learned from other physicians could potentially result in changes that affect disparities in care. Existing scholarship has documented institutional-level mechanisms such as socialization into a professional culture with norms and practices affecting patient care (e.g., Becker et al. 1961; DelVecchio Good et al. 2003; Merton 1957).

Although we can confidently rule-out statistical discrimination, our findings are merely consistent with the other two explanations above. Additional careful research is needed to further support or eliminate these or other mechanisms. One possible approach is to assess medical students' cognitive biases, both implicit and explicit (e.g., Sabin, Rivara and Greenwald 2008) over time as they progress through medical education. Answering the question of whether disparities in service delivery to clients by professionals derive from learned behaviors acquired during professional training or manifestations of existing biases is critical for addressing the pervasive and enduring disparities affecting many professions.

REFERENCES

Agency for Healthcare Research and Quality. 2008. *2007 National Healthcare Disparities Report*. Agency for Healthcare Research and Quality: Rockville, MD. Retrieved March 30, 2008. (http://www.ahrq.gov/qual/nhdr07/nhdr07.pdf).

Aigner, D. J., G. G. Cain. 1977. Statistical theory of discrimination in labor markets. *Indust. Labor Relations Rev.* **30**(2):175–187.

Altonji, J. G., C. R. Pierret. 2001. Employer learning and statistical discrimination. *Quart. J. Econom.* **116**:313-50.

Arrow, K. J. 1972. Some Mathematical Models of Race Discrimination in the Labor Market. Pp. 187-204 in *Racial Discrimination in Economic Life,* edited by A.H. Pascal. Lexington MA: D.C. Heath.

Ayres, I., P. Siegelman. 1995. Race and Gender Discrimination in Bargaining for a New Car. *Amer. Econom. Rev.* **85**(3): 304-321.

Balsa, A. I., T. G. McGuire. 2001. Statistical discrimination in health care. *Journal of Health Economics*, **20**(6): 881-907.

_____. 2003. Prejudice, clinical uncertainty and stereotyping as sources of health disparities. *J. Health Econom.* **22**(1): 89-116.

Barrows, H. S. 1971. *Simulated patients (programmed patients) the development and use of a new technique in medical education*. Springfield, IL: Charles C Thomas.

_____. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Academic Med.* **68**(6): 443–453.

Baumle, A. K., M. Fossett. 2005. Statistical Discrimination in Employment: Its Practice, Conceptualization, and Implications for Public Policy. *Amer. Behavioral Scientist.* **48**(9): 1250-1274.

Beach, M. C., M. Rosner, L. A. Cooper, P. S. Duggan, J. Shatzer. 2007. Can patient-centered attitudes reduce racial and ethnic disparities in care? *Academic Med.* **82**, 193-198.

Becker, G. S. 1971. *The Economics of Discrimination, 2nd Edition*. Chicago, IL: Chicago University Press.

Becker, H. S., B. Geer, A. L. Strauss, E. C. Hughes. 1961. *Boys in white: Student culture in medical school.* Chicago IL: University of Chicago Press.

Bertrand, M., S. Mullainathan. 2004. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *Amer. Econom. Rev*. **94**(4): 991-1013.

Betancourt, J. R. 2006. Eliminating Racial and Ethnic Disparities in Health Care: What Is the Role of Academic Medicine? *Academic Med.* **81**:788–792.

Chandra, A., D. O. Staiger. 2010. Identifying Provider Prejudice in Healthcare. NBER Working Paper No. 16382.

Chen, J., S. S. Rathore, M. J. Radford, Y. Wang, H. M. Krumholz. 2001. Racial differences in the use of cardiac catheterization after acute myocardial infarction. *New England J. Med.* **344**: 1443-1449.

Chin, M. H., C. A. Humikowski. 2002. When Is Risk Stratification by Race or Ethnicity Justified in Medical Care? *Academic Med.* **77**(3):202-208.

Colliver, J. A., M. H. Swartz, R. S. Robbs. 2001. The effect of examinee and patient ethnicity in clinical-skills assessment with standardized patients. *Adv. Health Sci. Ed.* **6**: 5-13.

Cooke, M., D. M. Irby, W. Sullivan, K.. M. Ludmerer. 2006. American Medical Education 100 Years after the Flexner Report. *New England J. Med.* **355**: 1339-1344.

Cooper, L. A., D. L. Roter, R. L. Johnson, D. E. Ford, D. M. Steinwachs, N. R. Powe. 2003. Patient-Centered Communication, Ratings of Care, and Concordance of Patient and Physician Race. *Ann Intern. Med.* **139**:907-915.

Cooper-Patrick, L., J. J. Gallo, J. J. Gonzales, H. T. Vu, N. R. Powe, C. Nelson, D. E. Ford. 1999. Race, Gender, and Partnership in the Patient-Physician Relationship. *JAMA.* **282**: 583-589.

Correll, S. J., S. Benard. 2006. Biased estimators? Comparing status and statistical theories of gender discrimination. *Soc. Psych. Workplace.* **23**: 89-116.

Cunningham, W. A., K. J. Preacher, M. R. Banaji. 2001. Implicit Attitude Measures: Consistency, Stability, and Convergent Validity. *Psych. Sci.* **12**(2): 163-170.

DelVecchio Good, M. J., C. James, B. J Good, A. E Becker. 2003. The culture of medicine and racial, ethnic, and class disparities in healthcare. Pp. 594-625 in *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care*, edited by B.D. Smedley, A.Y. Stith, and A.R. Nelson. Washington, DC: National Academies Press.

Devi, S. 2008. U.S. health care still failing ethnic minorities. *Lancet.* **371**: 1903-1904.

England, P. 1994. Neoclassical Economists' Theories of Discrimination. Pp. 59-69 in *Equal employment opportunity: Labor market discrimination and public policy*, edited by P. Burstein. Hawthorne, NY: Aldine de Gruyter.

Fennell, M. L. 2005. Racial Disparities in Care: Looking Beyond the Clinical Encounter. *Health Services Res.* **40**(6): 1713-1721.

Fix, M., M. A. Turner, Eds. 1998. *A national report card on discrimination in America: The role of testing.* Washington, DC: Urban Institute Press.

Green, A. R., D. R. Carney, D. J. Pallin, L. H. Ngo, K. L. Raymond, L. Iezzoni, M. R. Banaji. 2007. Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *J. General Internal Med.* **22**(9): 1231-1238.

Gross, C. P., B. D. Smith, E. Wolf, M. Andersen. 2008. Racial disparities in cancer therapy : did the gap narrow between 1992 and 2002? *Cancer.* **112**:900-8.

Heckman, J.J., P. Siegelman. 1993. The Urban Institute Audit Studies: Their Methods and Findings: Response to Comments by John Yinger. In M. Fix and R. J. Struyk, eds.: *Clear and Convincing Evidence: Measurement of Discrimination in America.* Washington, D.C.: Urban Institute Press; distributed by University Press of America: Lanham, MD: 271-75.

Heritage, J., D. W. Maynard. 2006. Problems and prospects in the study of physician-patient interaction: 30 years of research. *Annual Rev. Sociology*. **32**:351–74.

Institute of Medicine. 2003. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care.* Washington, DC: The National Academies Press.

Klonoff, E. A. 2009. Disparities in the provision of medical care: an outcome in search of an explanation. *J. Behavioral Med.* **32**:48–63.

Knowles, J., N. Persico, P. Todd. 2001. Racial Bias in Motor Vehicle Searches: Theory and Evidence. *J. Political Econom.* **109**(1): 203-229.

Ladd, H. F. 1998. Evidence on Discrimination in Mortgage Lending. *J. Econom. Perspectives.* **12**(2): 41-62.

LaVeist, T. A., A. Nuru-Jeter. 2002. Is Doctor-Patient Race Concordance Associated with Greater Satisfaction with Care? *J. Health Soc. Behavior.* **43**(3): 296-306.

Liaison Committee on Medical Education 2008. *Functions and Structure of a Medical School: Standards for Accreditation of Medical Education Programs Leading to the M.D. Degree*. Washington DC: LCME Secretariat Association of American Medical Colleges.

Lee, J. 2000. The Salience of Race in Everyday Life: Black Customers' Shopping Experiences in Black and White Neighborhoods. *Work and Occupations*. **27**(3): 353-376.

Lutfey, K. E., J. D. Ketcham. 2005. Patient and Provider Assessments of Adherence and the Sources of Disparities: Evidence from Diabetes Care. *Health Services Res*. **40**(1): 2–25.

McConnell, A. R., J. M. Leibold. 2001. Relations among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes. *J. Experiment. Soc. Psych.* **37**: 435– 442.

McGuire, T. G., J. Z. Ayanian, D. E. Ford, R. E. M. Henke, K. M. Rost A. M. Zaslavsky. 2008. Testing for Statistical Discrimination by Race/Ethnicity in Panel Data for Depression Treatment in Primary Care. *Health Services Res.* **43**(2): 531-551.

Merton, R. K. 1957. Some Preliminaries to a Sociology of Medical Education. Pp. 3-79 in *The Student Physician: Introductory Studies in the Sociology of Medical Education,* edited by R.K. Merton, G.G. Reader, and P.L. Kendall. Cambridge, MA: Harvard University Press.

_____. 1987. Three fragments from a sociologist's notebooks: Establishing the phenomenon, specified ignorance, and strategic research materials. *Annual Rev. Sociology.* **13**: 1-29.

Mizrahi, T. 1986. *Getting Rid of Patients.* New Brunswick, NJ: Rutgers University Press.

National Partnership for Action. 2010. *Changing Outcomes – Achieving Health Equity, The National Plan for Action.* National Partnership for Action (NPA) to End Health Disparities, Office of Minority Health, U.S. Department of Health and Human Services. Available online at: http://www.minorityhealth.hhs.gov/npa/images/plan/nationalplan.pdf.

National Research Council. 2004. *Measuring Racial Discrimination*. Washington DC: National Academies Press.

Neumark, D., R. J. Bank, K. D. Van Nort. 1996. Sex Discrimination in Restaurant Hiring: An Audit Study. *Quart. J. Econom.* **111** (3):915-42.

Nickell, S. 1981. Biases in Dynamic Models with Fixed Effects. *Econometrica*, **49**(6): 1417-1426.

Orsi, J. M., H. Margellos-Anast, S. Whitman. 2010. Black–White Health Disparities in the United States and Chicago: A 15-Year Progress Analysis. *Amer. J. Public Health.* **100**(2): 349-56.

Pager, D. 2003. The Mark of a Criminal Record. *Amer. J. Sociology.* 108(5): 937–75.

Pager, D., D. Karafin. 2009. Bayesian Bigot? Statistical Discrimination, Stereotypes, and Employer Decision Making. *Ann. Amer. Acad. Political Soc. Sci.* **621**(1): 70-93.

Pager, D., H. Shepard. 2008. The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets. *Annual Rev. Sociology.* **34**:181–209.

Phelps, E. S. 1972. The Statistical Theory of Racism and Sexism. *Amer. Econom. Rev.* 62(4): 659-661.

Pletcher, M. J., S. G. Kertesz, M. A. Kohn R. Gonzales. 2008. Trends in Opioid Prescribing by Race/Ethnicity for Patients Seeking Care in US Emergency Departments. *JAMA.* **299**(1):70-78.

Price, J., J. Wolfers. 2007. Racial Discrimination Among NBA Referees. Working Paper, National Bureau of Economic Research, Cambridge, MA.

Quillian, L. 2006. New Approaches to Understanding Racial Prejudice and Discrimination. *Annual Rev. Sociology.* **32**:299–328.

Richeson, J. A., J. N. Shelton. 2005. Brief Report: Thin Slices of Racial Bias. *J Nonverbal Behav.* **29**(1): 75-86.

Ross, S. L., M. A. Turner. 2005. Housing Discrimination in Metropolitan America: Explaining Changes between 1989 and 2000. *Soc. Problems*. **52**(2): 152–180.

Russo, J. E., M. G. Meloy, T. J. Wilks. 2000. Predecisional Distortion of Information by Auditors and Salespersons. *Management Sci*. **46**(1):13-27.

Sabin, J. A., F. P. Rivara A. G. Greenwald. 2008. Physician implicit attitudes and stereotypes about race and quality of medical care. *Medical Care.* **46**:678-685.

Schnittker, J., K. Liang. 2006. The Promise and Limits of Racial/Ethnic Concordance in Physician-Patient Interaction. *J. Health Politics, Policy and Law*, **31**(4):811-838.

Schulman, K. A., J. A. Berlin, W. Harless, J. F. Kerner, S. Sistrunk, B. J. Gersh, R. Dubé, C. K. Taleghani, J. E. Burke, S. Williams, J. M. Eisenberg, J. J. Escarce. 1999. The Effect of Race and Sex on Physicians' Recommendations for Cardiac Catheterization. *New England J. Med*. **340**:618-26.

Shrout, P. E., J. L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psych. Bull.* **86**(2): 420-428.

Simons, T., R. Friedman, L. A. Liu, J. McLean Parks. 2007. Racial Differences in Sensitivity to Behavioral Integrity: Attitudinal Consequences, In-Group Effects, and "Trickle Down" Among Black and Non-Black Employees. *J. Appl. Psych.* 92(3): 650–665.

Siniver, E. 2011. Testing for Statistical Discrimination: The Case of Immigrant Physicians in Israel. *Labour*. **25**(2):155-166.

Smith, R. C. 2003. An Evidence-Based Infrastructure for Patient-Centered Interviewing. Pp. 148-163 in *The Biopsychosocial Approach: Past, Present, and Future,* edited by R.M. Frankel, T.E. Quill, and S.H. McDaniel. Rochester NY: University of Rochester.

Spiro, H. 1992. What Is Empathy and Can It Be Taught? *Ann. Internal Med.* **116**(10):843-846.

Stainback, K., D. Tomaskovic-Devey, S. Skaggs. 2010. Organizational Approaches to Inequality: Inertia, Relative Power, and Environments. *Annu. Rev. Sociol.* **36**:225–47.

Tetlock, P. E., G. Mitchell. 2008. Calibrating Prejudice in Milliseconds. *Soc. Psych. Quart.* **71**(1): 12–16.

_____. 2009. Implicit Bias and Accountability Systems: What Must Organizations Do to Prevent Discrimination? *Res.Organ. Behavior*. **29**: 3–38.

Tomaskovic-Devey, D., S. Skaggs. 1999. An Establishment-Level Test of the Statistical Discrimination Hypothesis. *Work and Occupations.* **26**: 422-445.

Vaccarino, V., S. S. Rathore, N. K. Wenger, P. D. Frederick, J. L. Abramson, H. V. Barron, A. Manhapra, S. Mallik, H. M. Krumholz, the National Registry of Myocardial Infarction Investigators. 2005. "Sex and Racial Differences in the Management of Acute Myocardial Infarction, 1994 through 2002." *New England J. Med.* **353**:671-682.

Van Ryn, M. 2002. "Research on the provider contribution to race/ethnicity disparities in medical care." *Medical Care.* **40**: I-140-I-151.

Yinger, J. 1986. Measuring Racial Discrimination with Fair Housing Audits: Caught in the Act. *Amer. Econom. Rev.* **76**(5):881-893.

_____. 1996. Discrimination in Mortgage Lending: A Literature Review. In John Goering and Ron Wienk, eds. *Mortgage Lending, Racial Discrimination, and Federal Policy*, Washington, D.C.: Urban Institute Press. Pp: 29-74.

Table 1: Observations by Cohort, Case and Standardized Patient Race, and Available Cohort Demographic Characteristics.

Case	SP Race	Class of 2009	Class of 2010	Class of 2011
Year 1: Hx	White	81	90	85
	Black	14	8	18
Year 2: HxPE	White	86	91	63
	Black	5	9	37
Cohort Demographics				
Percent Female		48%	39%	52%
Percent Black		NA	NA	14%
Percent White		NA	NA	65%

Table 2: Counts**,** means and standard deviations for SP encounter outcomes by year and SP race. The top panel reports the graded outcomes provided by OSM for all three medical student cohorts. The bottom panel reports outcomes from the video coding of the class of 2011 cohort encounters. White-Black differences in outcomes by year are given in the "Difference" column, along with the significance results of a t-test (based on the pooled standard errors of the differences) testing whether the Second Year White-Black difference is an increase relative to the First Year White-Black difference for each outcome.

	White SP			**Black SP**			**Difference** (White – Black)		
	N	Mean	*(SD)*	N	Mean	*(SD)*	Δ	*(SE)*	Significant Increase?
Graded Encounter Outcomes from All Three Cohorts									
History									
First Year	256	0.77	*(0.14)*	40	0.79	*(0.12)*	-0.02	*(0.02)*	ns
Second Year	240	0.84	*(0.13)*	51	0.85	*(0.09)*	-0.01	*(0.02)*	
Patient Satisfaction									
First Year	256	0.81	*(0.20)*	40	0.80	*(0.15)*	0.01	*(0.03)*	***
Second Year	240	0.80	*(0.22)*	51	0.65	*(0.20)*	0.15	*(0.03)*	
Patient-Physician Interaction									
First Year	256	0.85	*(0.13)*	40	0.87	*(0.10)*	-0.02	*(0.02)*	***
Second Year	240	0.84	*(0.13)*	51	0.72	*(0.14)*	0.13	***(0.02)***	
Coded Encounter Videos of the Class of 2011 Cohort									
Lean Towards SP									
First Year	83	3.55	*(1.37)*	18	3.62	*(1.31)*	-0.08	*(0.35)*	*
Second Year	62	4.83	*(1.59)*	37	4.11	*(1.49)*	0.72	*(0.32)*	
Smiling									
First Year	83	3.05	*(1.03)*	18	3.25	*(1.23)*	-0.21	*(0.28)*	*
Second Year	62	3.42	*(1.08)*	37	3.05	*(1.12)*	0.37	*(0.22)*	
Positive Demeanor									
First Year	83	4.70	*(1.02)*	18	4.91	*(0.94)*	-0.21	*(0.26)*	**
Second Year	62	5.19	*(0.87)*	37	4.82	*(0.79)*	0.37	***(0.17)***	

ns: not significant, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, two-tailed tests.

Table 3: Regressions estimating trends between training and disparities in care across three medical student cohorts using both linear SP controls and SP fixed effects models (46 SPs). Standard Errors appear in parentheses. Shaded results are for the key variable, *SPWHITE X YEAR*. N=587.

	History		**Patient Satisfaction**		**PPI**[a]	
	SP Controls	SP Fixed Effects	SP Controls	SP Fixed Effects	SP Controls	SP Fixed Effects
SPWHITE X YEAR	-0.005	-0.023	**0.123****	**0.107**†	**0.112*****	**0.101****
	(0.029)	*(0.038)*	*(0.047)*	*(0.059)*	*(0.029)*	*(0.036)*
YEAR	**0.072****	**0.088***	**-0.134****	**-0.126***	**-0.130*****	**-0.126*****
	(0.027)	*(0.035)*	*(0.043)*	*(0.055)*	*(0.029)*	*(0.034)*
SPWHITE	-0.022	—	0.001	—	-0.023	—
	(0.021)		*(0.035)*		*(0.021)*	
SPFEMALE	-0.020	—	-0.018	—	-0.018	—
	(0.015)		*(0.025)*		*(0.015)*	
SPAGE	**-0.008**†	—	-0.009	—	-0.005	—
(in days/3652.5)	*(0.004)*		*(0.007)*		*(0.004)*	
SPEXPERIENCE	0.003	-0.073	-0.015	-0.212	**-0.406***	-0.232
(in encounters/1000)	*(0.175)*	*(0.248)*	*(0.282)*	*(0.388)*	*(0.183)*	*(0.282)*
CLASS OF 2010	**-0.053*****	**-0.040**†	-0.004	0.016	**-0.035***	-0.024
	(0.015)	*(0.022)*	*(0.024)*	*(0.035)*	*(0.015)*	*(0.021)*
CLASS OF 2011	**-0.061*****	**-0.055****	**-0.066****	-0.023	**-0.130*****	**-0.106*****
	(0.014)	*(0.022)*	*(0.023)*	*(0.035)*	*(0.017)*	*(0.023)*
STUDENTFEMALE	**-0.025**†	**-0.029** †	0.029	0.012	-0.009	-0.011
	(0.015)	*(0.016)*	*(0.024)*	*(0.025)*	*(0.014)*	*(0.015)*
BOTHFEMALE	**0.037**†	**0.041**†	0.010	0.027	0.033	**0.039**†
	(0.022)	*(0.023)*	*(0.035)*	*(0.036)*	*(0.021)*	*(0.021)*
PPICHANGE	—	—	—	—	0.026	0.000
					(0.024)	*(0.030)*
CONSTANT	**0.882*****	**0.809*****	**0.869*****	**0.803*****	**0.971*****	**0.906*****
	(0.034)	*(0.017)*	*(0.055)*	*(0.026)*	*(0.033)*	*(0.016)*

† $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, two-tailed tests

[a] The PPI scores in this table were the scores provided by OSM. To ensure our results are robust to the PPI scoring change in the second year encounter for the class of 2011 cohort, we also converted the individual 14 item scores to binary scores by rounding prior to calculating the PPI. In each of the two models with this reconstructed PPI measure, the magnitude and significance of the *SPWHITE X YEAR* variable increased.

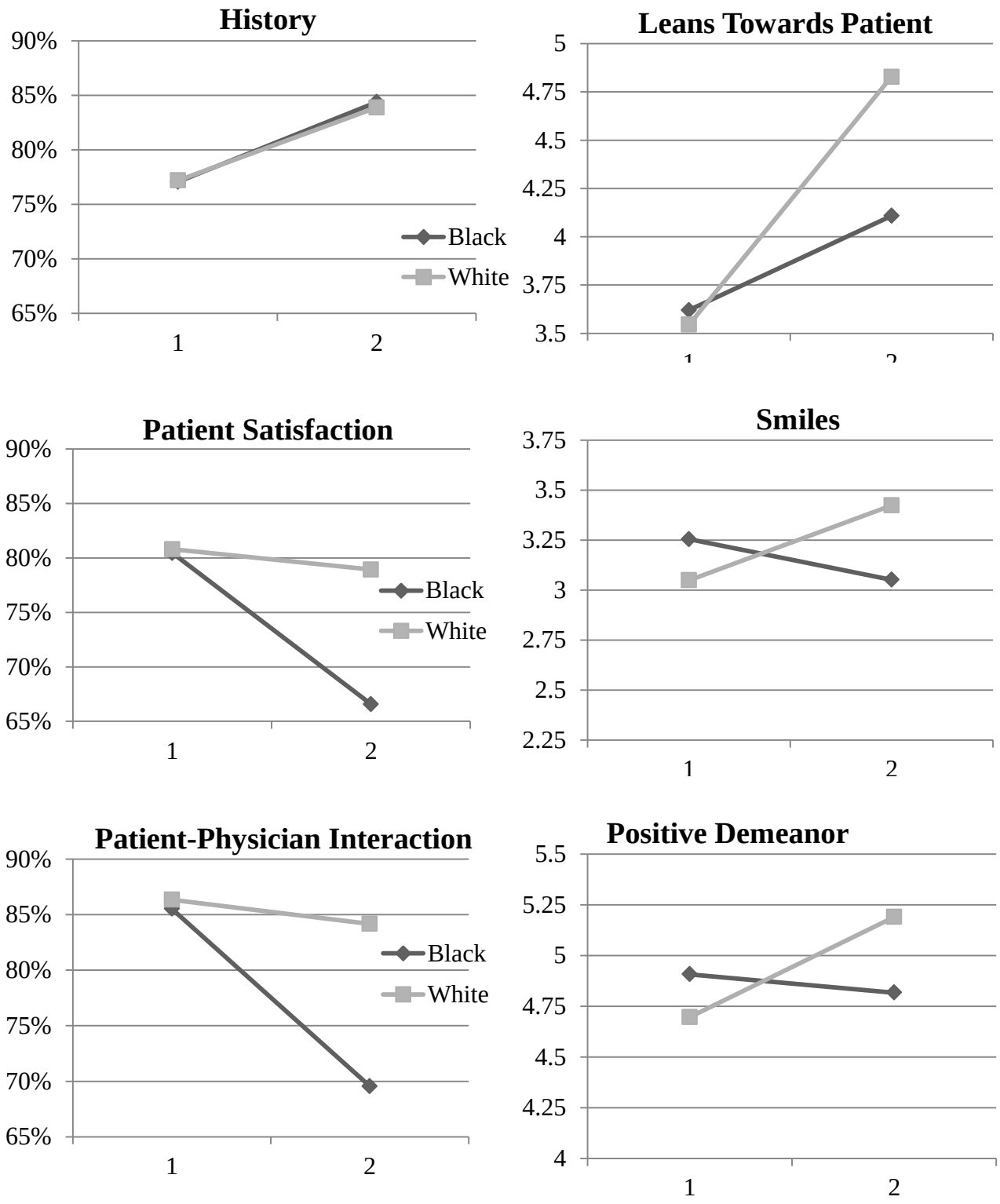
Table 4: Regressions of class of 2011 medical students' coded non-verbal behavior and demeanor and graded SP encounter outcomes on SP controls, student controls, and concordance effects. Standard Errors appear in parentheses. N=200.

	Non-Verbal Behaviors & Demeanor			**Graded Encounter Items**		
	Leans towards SP	Smiles	Positive Demeanor	History	Patient Satisfaction	PPI[a]
SPWHITE X YEAR	0.606 *(0.494)*	0.372 *(0.332)*	**0.614*** *(0.308)*	0.009 *(0.044)*	**0.123**† *(0.066)*	**0.174***** *(0.045)*
YEAR	0.191 *(0.432)*	-0.285 *(0.290)*	-0.032 *(0.269)*	0.056 *(0.039)*	**-0.150*** *(0.058)*	**-0.146***** *(0.039)*
SPWHITE	-0.167 *(0.619)*	-0.234 *(0.416)*	-0.418 *(0.386)*	-0.032 *(0.055)*	**0.151**† *(0.083)*	0.003 *(0.056)*
SPFEMALE	0.103 *(0.373)*	0.249 *(0.251)*	0.189 *(0.233)*	-0.035 *(0.033)*	-0.032 *(0.050)*	-0.052 *(0.034)*
SPAGE *(in days/3652.5)*	0.027 *(0.086)*	0.043 *(0.058)*	0.005 *(0.054)*	-0.003 *(0.008)*	-0.014 *(0.012)*	-0.002 *(0.008)*
SPEXPERIENCE *(encounters/1000)*	**7.55*** *(3.72)*	2.78 *(2.50)*	-2.67 *(2.32)*	0.138 *(0.330)*	0.036 *(0.495)*	-0.361 *(0.335)*
STUDENTWHITE	0.199 *(0.500)*	0.273 *(0.336)*	-0.008 *(0.312)*	0.028 *(0.045)*	**0.214**** *(0.067)*	**0.078**† *(0.045)*
STUDENTBLACK	0.090 *(0.417)*	0.246 *(0.280)*	0.205 *(0.260)*	0.008 *(0.037)*	**-0.132*** *(0.056)*	**-0.071**† *(0.038)*
STUDENTFEMALE	0.092 *(0.347)*	**0.683**** *(0.233)*	**0.455*** *(0.217)*	-0.041 *(0.031)*	0.009 *(0.046)*	-0.015 *(0.031)*
BOTHWHITE	0.576 *(0.633)*	0.126 *(0.425)*	0.276 *(0.395)*	0.012 *(0.057)*	**-0.140**† *(0.085)*	-0.025 *(0.057)*
BOTHBLACK	0.064 *(0.814)*	-0.056 *(0.547)*	-0.398 *(0.507)*	0.062 *(0.073)*	**0.241*** *(0.109)*	0.061 *(0.074)*
BOTHFEMALE	0.323 *(0.597)*	0.421 *(0.401)*	0.203 *(0.372)*	0.082 *(0.053)*	0.112 *(0.080)*	0.081 *(0.054)*
MATCHSEXANDRACE	**-1.058*** *(0.411)*	-0.205 *(0.276)*	-0.176 *(0.256)*	-0.040 *(0.037)*	**-0.129*** *(0.055)*	**-0.062**† *(0.037)*
CONSTANT	**3.251***** *(0.710)*	**2.319***** *(0.477)*	**4.621***** *(0.443)*	**0.793***** *(0.064)*	**0.686***** *(0.095)*	**0.802***** *(0.065)*

† $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[a] As before, when using a variant of the PPI score constructed by coercing the individual 14 items into binary scales by rounding, the magnitude and significance of the *SPWHITE X YEAR* variable increased relative to what is presented here.

Figure 1: Left Column: Predicted (from fixed effects models in Table 4) SP encounter outcomes by medical school year and SP race, based on data from all 3 cohorts, and using identical y-axis ranges. Right Column: Predicted (from models in Table 5) nonverbal behavior and demeanor outcomes plotted similarly for the class of 2011 cohort, and using 1.5-unit ranges of the 1-7 valued variables for the y-axes.



APPENDIX: Outcome Measure Instruments

HISTORY

The student asked the following (Please select Yes or No):

A. About any symptom I am having: (History of present illness)

1. Quality/Description?
2. Severity (How bad is it? Is it getting better? Worse? Unchanged?)
3. Timing (How long has this been going on? Is it constant?)
4. Context (Have I had these symptoms before?)
5. Modifying factors (Anything make it better? Worse? Have I tried anything for it - medications, other?)
6. Associated symptoms/signs (Symptoms/signs to ask about will depend on the specific chief complaint)
7. Impact on life (Any effect on my life?)

B. Past medical problems

8. About any past medical problems? and for each diagnosis:
9. When the diagnosis was made?
10. How the diagnosis was made?
11. How the medical problem has been treated (medications, other therapy) and/or is currently being treated?
12. About any complications/other related problems I have had?
13. If/how the diagnosis has affected my daily functioning?

C. Past surgical problems

14. About my surgical problems? and for each surgery:
15. When I had the surgery?
16. Why I had the surgery?
17. About any complications/other problems related to the surgery?

D. Medications

18. About my medications:
19. The dose and frequency of each medication?
20. Do I take any over-the-counter medications?
21. Do I take any other supplements?

E. Allergies

22. About my allergies? and for each allergy:
23. What was the specific allergic reaction?

F. Social and occupational history

24. My current occupation?
25. My marital status/living situation?
26. Do I feel safe at home?
27. Do I smoke, and if so, how much/how long?
28. Do I drink, and if so, how much/how often?
29. If I currently use or have used any illicit drugs and if so, which drugs?
30. About my diet and exercise habits?

G. Family history

31. Do any medical conditions run through my family?
32. Does heart disease run through my family, and if so, in whom?
33. Does diabetes run through my family and if so, in whom?
34. Does high blood pressure run through my family and if so, in whom?
35. Does cancer run through my family and if so, what cancer/in whom/at what age?

H. Sexual history

36. Am I currently sexually active?
37. Are my partners' male, female, or both?
38. How many sexual partners have I had in the past year?
39. Have I ever had a sexually transmitted disease and if so, was I treated?
40. Do I/does my partner use condoms?
41. Do I/does my partner use any form of contraception?
42. Have I ever been tested for HIV?

I. Review of Systems (ROS)

43. Began by explaining to me what he/she means by ROS (i.e. that these are screening questions)?
44. Had an organized sequence in asking me the ROS questions?

Asked me about whether I have any 2 items in each of the following categories:

45. General/systemic - Fever, chills, night sweats, any changes in weight, any changes in appetite, fatigue
46. Skin/Integument - Rashes, lumps, itchiness, changes in hair or nails
47. Neurologic/Psychiatric - Headaches, weakness, numbness, change in memory, difficulties with speech, difficulties walking; depressed mood, excessive moodiness, nervousness, difficulty sleeping
48. HEENT - Changes/problems with vision (nearsightedness, blurred vision, double vision, spots), changes/problems with hearing, ringing in ears, dizziness, lumps/swollen glands in neck
49. Endocrine - Feeling hot or cold at temperatures where others are comfortable, excessive hunger, excessive thirst, frequent urination
50. Breasts - Lumps, pain, nipple discharge, skin changes
51. Cardiovascular - Chest pain, heart racing/fluttering, shortness of breath with activity, awakening at night b/c of shortness of breath, sleeping on more than one pillow b/c of shortness of breath, near-blackouts or blackouts, swelling in legs, pain in calf with walking which is relieved with rest
52. Pulmonary - Cough, wheezing, shortness of breath
53. Gastrointestinal - Difficulty swallowing, heartburn, nausea, vomiting, abdominal pain, constipation, diarrhea, changes in stool size, black/tarry stool, blood in stool
54. Genitourinary - Burning/pain with urination, blood in urine, need to urinate urgently/suddenly, getting up in the middle of the night to urinate, loss of control of urination, difficulty getting urine stream started, genital discharge, genital sores
55. Musculoskeletal - Muscle weakness or pain, joint pain or swelling, back pain, limitations in movement or activity
56. Hematologic - Abnormal/excessive bleeding, easy bruising, enlarged lymph nodes in neck/armpits/groin

J. OB/GYN history (Female only):

57. How old I was when I first had my period (menarche)?
58. Have I ever been pregnant?
59. If I have been pregnant: # of deliveries, about any complications related to pregnancy?
60. Do I currently have periods?
61. Are my periods regular and how long do they last?
62. Do I have heavy bleeding or any other problems related to my periods?
63. When was my last period?

PATIENT PHYSICIAN INTERACTION

During the encounter the student did the following (select Yes or No):

1. Greeted me, introduced himself/herself?
2. Called me by name?
3. Used appropriate eye contact?
4. Showed interest/respect for me throughout the interview? (Open body language listened carefully, appropriate facial expressions and tone of voice)
5. Used language that I could understand? (avoided technical terms)
6. Started with open-ended questions?
7. Progressed with specific questions?
8. Avoided presumptive/leading questions?
9. Allowed me to speak without interruption?
10. Checked to make sure that he/she understood what I was saying?
11. Was organized in the order that he/she asked me questions?
12. Summarized the information that he/she gathered?
13. Checked to make sure that I understood what he/she was saying?
14. Closed the encounter by telling me his/her initial impression of what was going on and described what he/she thought needed to be done?

PATIENT SATISFACTION

Based on your interaction with the student, please select the best response:

1. I would come back to see this student-doctor. {definitely not, probably not, might, probably would, definitely would}
2. I would recommend this student-doctor to a relative or friend {definitely not, probably not, might, probably would, definitely would}

NONVERBAL / DEMANOR CODING ITEMS

For each of the following, score on a 1-7 scale. 1=All the time, 4=About half the time, 7=Never.

1. How frequently did the medical student lean towards the patient?
2. How frequently did the medical student smile during the interaction?

For each of the following qualities, rate the degree to which the medical student appeared to exhibit these qualities during their interaction with the patient. Medical student seemed (1-7, 1=extremely, 4=Somewhat, 7=not at all):

3. Likeable
4. Warm
5. Friendly
6. Pleasant