Does Affirmative Action Lead to Mismatch? A New Test and Evidence[*]

Peter Arcidiacono Esteban M. Aucejo Hanming Fang[†]

October 14, 2008

Abstract

We argue that once we take into account the students' rational enrollment decisions, mismatch in the sense that the intended beneficiary of affirmative action admission policies are made worse off could occur only if selective universities possess private information about students' post-enrollment treatment effects. This necessary condition for mismatch provides the basis for a new test. We propose an empirical methodology to test for private information in such a setting. The test is implemented using data from Campus Life and Learning Project (CLL) at Duke. Preliminary evidence shows that Duke does possess private information that is a statistically significant predictor of the students' post-enrollment academic performance, but Duke's private information only explains a very small percentage of the variation in student performance. We also propose strategies to evaluate more conclusively whether the evidence of Duke private information has generated mismatch.

Keywords: Mismatch; Private information; Affirmative Action

JEL Classification Codes: D8, I28, J15.

[*] Preliminary and incomplete. All comments are welcome. We would like to thank Judy Chevalier, Jon Levin, Jim Levinsohn, Jeff Smith, Justin Wolfers and seminar participants at Wharton School for helpful comments. We are grateful to Ken Spenner for giving us access to the Duke Campus Life and Learning Survey data. All remaining errors are ours. All remaining errors are ours.

[†] Department of Economics, Duke University, 213 Social Sciences Building, P.O. Box 90097, Durham, NC 27708-0097. Arcidiacono and Fang are also affiliated with NBER. Email addresses are respectively: peter.arcidiacono@duke.edu, esteban.aucejo@duke.edu, hanming.fang@duke.edu

1 Introduction

The use of racial preferences in college and university admissions has generated much debate. Proponents of racial preferences argue that race-conscious admissions are important both for helping minorities overcome the legacy of the institutionalized discrimination and for majority students to receive the benefits from diverse classrooms.[1] Opponents of racial preferences assert that race-conscious admissions are unfair and may actually be damaging to the intended beneficiaries by placing them at institutions where they are unlikely to succeed.[2]

Recently the controversy over race-conscious admission policies has increasingly moved from a normative to a positive perspective. On one front, several papers attempted to empirically examine the educational benefits of attending racially diverse colleges. For example, Black, Daniels and Smith (2001) found a positive relationship between proportion of blacks in the college attended and the post-graduate earnings in the National Longitudinal Survey of Youth; Arcidiacono and Vigdor (2007), using data on graduates of 30 selective universities in College and Beyond data, found only weak evidence of any relationship between collegiate racial composition and the post-graduation outcomes of white or Asian students. Duncan et. al. (2006), exploiting conditionally random roommate assignment at one large public university, found that cross-racial exposure influences individual attitudes and friendship patterns.

A second front, spurred by the provocative article of Sander (2004) and followed up by Ayres and Brooks (2005), Ho (2005), Chambers et. al. (2005), Barnes (2007) and Rothstein and Yoon (2008), attempts to empirically examine whether the effects of affirmative action policies on the intended beneficiaries is positive or negative. These papers essentially tests for the so-called "*mismatch hypothesis*," i.e. whether the outcome of minority students might have been worsened as a result of attending a selective university relative to attending a less selective school.

The above papers all attempt to test the mismatch hypothesis by comparing the "outcome" (e.g., GPA, bar passage, post-graduate earnings etc.) of the minority students enrolled in elite universities relative to the corresponding *counterfactual* outcome when these minority students attend less selective universities. As well summarized in Rothstein and Yoon (2008), the papers differ in how the counterfactual outcomes are assessed. For example, Sander (2004) first used a comparison of black and white students with the same *observable* credentials, who typically attend

[1] In both *Regents of University of California v. Bakke* 438 U.S. 265 (1978) and more recently in *Grutter v. Bollinger*, 539 U.S. 306 (2003), the Supreme Court ruled that the educational benefits of a diverse student body is a compelling state interest that can justify using race in university admissions.

[2] See Kellough (2006) for a concise and up-to-date introduction to various arguments for and against affirmative action.

different law schools because of affirmative action, to estimate a negative effect of selectivity on law school grades; he then included both selectivity and grades in a regression for graduation and bar passage where he found that both selectivity and grades have positive coefficient, with the latter much larger than the former. Combining these two findings, he concluded that, on net, preferences in law school admission in favor of black students depressed black outcomes because such preferences led black students into more selective schools, lowering their law school grades, which swamps the positive effective of attending a selective school on their graduation and passing the bar. Ayres and Brooks (2005), Ho (2005), Chambers et. al. (2005) and Barnes (2007), however, used versions of *selective-nonselective comparison,* i.e., comparing students of the same race and same observable admission credentials who attend more- and less-selective schools to assess whether attending more selective schools has negative effects.[3] All strategies used above to assess the counterfactual outcome are likely to yield biased estimates when there are *unobservable* characteristics that may be considered in admission but unobserved by researchers. For example, the selective-unselective comparison used by Ayres and Brooks (2005), Ho (2005), Chambers et. al. (2005) and Barnes (2007) are likely to underestimate mismatch effect because those who are admitted to more selective schools are likely to have better unobserved credentials.[4] In contrast, Sanders (2004), by attributing the black's lower grades in selective schools to school selectivity instead of potential unobserved credentials, is likely to overstate the mismatch effect. Finally, Rothstein and Yoon (2008) used both the selective-unselective and the black-white comparisons to provides bounds for the mismatch effect in law school. They find no evidence of mismatch effects on any students' employment outcomes or on the graduation or bar passage rates of black students with moderate or strong entering credentials. However, they did find mismatch effect for less-qualified black students who typically attend second- or third-tier schools. To summarize, the existing literature on the mismatch effect differs in the empirical strategy used to assess the counterfactual outcome when minority students attend less selective universities; and the evidence

[3] Barnes (2007) also explains that the performance for black students may suffer in a selective school both because of mismatch, i.e., they are over-placed in such selective schools, or because there are race-based barriers to effective learning in selective schools.

[4] Dale and Krueger (2002) proposed and applied a strategy to control for the unobservable credentials in estimating the treatment effect of attending highly selective colleges by comparing students attending highly selective colleges with others admitted to these schools but enrolled elsewhere. Ayres and brooks (2005) and Sanders (2005b) also attempted to approximately apply the Dale and Krueger strategy by comparing law students who reported attending their first choice schools with those who reported attending their second choices because their first choices were too expensive or too far from home. The problem obviously is that they do not know whether those reporting attending their second choice would have been admitted to the schools attended by the former group, thus it is not at all clear that such a strategy does anything to control for unobserved credentials.

is mixed.

Besides the difficulties in evaluating the counterfactual outcomes, the existing literature also faces other challenges. To illustrate this point, suppose that one can convincingly establish that blacks are less likely to pass bar exams after attending an elite law school. Does this necessarily mean that blacks are worse off in an *ex ante* expected utility sense? If attending an elite university also makes it possible for blacks to be high-profile judges, and if the outcome of being a high-profile judge is valued by blacks much higher than passing the bar exam, blacks could still be better off *ex ante* under affirmative action. Alternatively, it is possible that elite universities may provide amenities to minority students that more than compensate the worse outcome measures that are examined by the researcher, thus making the minority students better off *ex ante* in an expected utility sense.

In this paper we take a new and complementary viewpoint to the above-mentioned literature on mismatch by bringing to the center the *rational decision* of the minority students who are offered admission to a selective school, possibly due to affirmative action policies. The question we ask is, why would students be willing to enroll themselves at schools where they cannot succeed, as the mismatch hypothesis stipulates? Posing the question in this way immediately leads us to focus our attention to the role of asymmetric information. In Section 2 we show that a *necessary condition* for mismatch to occur once we take into account the minority students' rational enrollment decisions is that the selective university has private information about the treatment effect of the students. In the absence of asymmetric information about her treatment effect in the selective university (relative to attending a non-selective university), a minority student will choose to enroll in the selective university only if her treatment effect is positive, thus there is no room for mismatch to occur; however, when the selective university has private information about a minority student's treatment effect, it is possible that a minority student with a negative treatment effect may end up enrolling in the selective university if offered admission. The reason is simple: when the minority student decides whether to enroll in the selective university, she can only condition her decision on the event that her treatment is above its admission threshold; when the selective university's admission threshold for the minority student is negative, due to its desire to satisfy a diversity constraint for example, it may still be optimal for a minority student with a negative treatment effect to enroll as long as the average treatment effect conditional on admission is higher than that from the non-selective university.

The central message from the simple model in Section 2 is that the presence of private information by the selective university regarding the students' treatment effect is a necessary condition for mismatch effect as a result of affirmative action. This statement is true from a purely expected utility perspective, and thus is not invalidated by the challenges we mentioned previously for the existing literature. This simple observation leads to a novel test for a necessary condition for mismatch, which is a test for whether selective universities possess private information regarding the students they admit. We will emphasize that our test is only a test for *necessary condition*: if we find strong evidence for asymmetric information, it does not necessarily imply that mismatch has occurred; but if we find no evidence for asymmetric information, then we can rule out mismatch without having to rely on strong unverifiable assumptions needed for the assessment of counterfactual outcomes. We will also discuss in Section 6 how we can follow up our necessary condition test with additional data collection to more conclusively establish the presence or absence of mismatch. It is also important to note that, regardless of whether we can empirically establish the presence/absence of mismatch, our simple theory highlighting the rational enrollment decisions of the students naturally suggests policies that will be effective to decrease the possibility of mismatch, namely, to increase the information flow from the selective university to the minority students that can assist them in predicting their post-enrollment educational outcomes.

In Section 3 we propose a non-parametric method to test for asymmetric information. We assume that the researcher has access to the elite university's assessment of the applicants, the applicants' subjective expectation about their post-enrollment performance in the selective university and their actual performance. We show that the celebrated Kotlarski (1967) theorem can be used to decompose the private information possessed by the applicant, the private information possessed by the selective university, and the information common to the selective university and the applicant but unobserved to the researcher.[5] We propose an estimation method after the Kotlarski decomposition to test whether the selective university possess private information important for the prediction of the students' actual post-enrollment outcomes.

In Sections 4 and 5, we apply our test for private information using data from the Campus Life and Learning Project (CLL), which surveys two recent consecutive cohorts of Duke University students before and during college. The survey was completed by 1181 randomly selected students providing information about college expectations, social and family background, satisfaction measures and provides confidential access to students information records. The key features of the data for our purposes is that we have Duke Admission Office's ranking of the applicants as well as the student's pre-enrollment expectations about their grade point average. We also have a rich set of control variables about the students' family and high school background. The information contained in CLL conforms to the data required for implementing our empirical strategy described

[5] Kotlarski theorem has been applied in economics in Krasnokutskaya (2004) and Cunha, Heckman and Navarro (2005).

in Section 3.

We test whether Duke's private information is important to outcomes such as grade point average and graduation after conditioning on what is in the student's information set, including the private information in the student's expected grade point average. Not only is Duke's private information important for both grades and graduation rates even after conditioning on the student's information set, but we also find that the student has virtually no private information on their probabilities of succeeding. That is, once we condition on Duke's information set, the student's expected grade point average is virtually uncorrelated with their grades and their probability of graduating.

These results are based upon outcomes but our line of argument is on utility. In order for these results to be informative to the mismatch debate, it must be the case that students value success. We use information from subjective satisfaction measures in the CLL to establish the link between success in school and utility. While these tests suggest that Duke's private information is important predictor of student success and that success is related to student satisfaction, this is only a necessary condition for mismatch to hold. Students may still have found Duke to be the utility-maximizing choice even if they had known Duke's private information. We describe some ways of establishing whether or not mismatch is occurring in the conclusion.

The remainder of the paper is structured as follows. In Section 2 we present a simple model of a selective university's admission problem with rational students to clarify the key concepts of mismatch in our framework, and illustrate that the selective university's private information is a necessary condition for mismatch to occur; in Section 3 we propose a semi-parametric estimation method to test for private information; Section 4 describes the Campus Life and Learning (CLL) Project data that we use in our application to test for private information; Section 5 presents our preliminary results; Section 6 discusses two potential avenues to provide more conclusive evidence for mismatch; and Section 7 discusses and concludes.

2 The Model

Consider two universities that differ in selectiveness. For convenience, suppose that only one university is selective. We refer to the selective university as the elite university. The elite university has a capacity of C; but the non-selective university, which essentially encompasses all the other options for the students in our model, does not have a capacity constraint.

The students belong to one of two racial groups, and for concreteness, we will call them "White (w)" and "Black (b)." The total number of race r applicants is given by N_r for $r \in \{w, b\}$. Let

$T_r \in R$ denote the "treatment effect" of a student with race $r \in \{w, b\}$ from attending the elite university. The "treatment effect" measures the difference in a student's outcome from attending the elite university instead of her second option (which in this model is the non-elite university). Importantly, this treatment effect is determined by the quality of matching between the student's own characteristics and the universities. To the extent that the non-elite university is better suited to some students, T_r could be negative. In the population of race r students, T_r is distributed according to a continuous CDF F_r with density function f_r.

We assume that the objective of the elite university is to maximize the total treatment effect for the admitted students subject to the capacity constraints, and if appropriate, the diversity constraint as well. This is a useful starting point. Alternatively, the elite university may want to maximize the total outcomes for its students. We assume that the student is risk neutral, and thus will choose a university (if she is admitted) that offers her higher treatment effect.

2.1 The Elite University's Problems

2.1.1 The First Benchmark: Symmetric Information and No Diversity Concerns

We first consider the benchmark case where each student knows about her own treatment effect and the elite university does not have any diversity constraint. That is, the elite university does not have any private information about the student's treatment effect. In this symmetric information case, it is clear that any student will enroll in the elite university if she knows that her treatment effect is positive.

Under the assumption that the elite university's objective is to maximize the total treatment subject to an enrollment capacity constraint, its problem can be written as:

$$\max_{\{T_r^* \geq 0\}} \sum_{r \in \{w,b\}} N_r \int_{T_r^*}^{\infty} T_r f_r\left(T_r\right) dT_r \tag{1}$$

$$s.t. \quad \sum_{r \in \{w,b\}} N_r \left[1 - F_r\left(T_r^*\right)\right] \leq C. \tag{2}$$

First note that in any optimum the elite university will choose admission thresholds exceeding zero, that is, $T_r^* \geq 0$ will always be satisfied at optimum. Let $\mu \geq 0$ denote the multiplier associated with the capacity constraint. Taking the first order condition with respect to $T_r^*, r \in \{w, b\}$, we obtain:

$$\begin{aligned} -N_r T_r^* f_r\left(T_r^*\right) + \mu N_r f_r\left(T_r^*\right) &= 0 \text{ if } \mu > 0; && (3) \\ N_r T_r^* f_r\left(T_r^*\right) &= 0 \text{ if } \mu = 0. && (4) \end{aligned}$$

Thus, if the enrollment capacity constraint for the elite university is binding, i.e., if $\mu > 0$, it immediately follows from (3) that $T_r^* = \mu$ for both $r \in \{w, b\}$. Thus $T_w^* = T_b^* = T^*$ where T^* uniquely solves

$$\sum_{r\in\{w,b\}} N_r \left[1 - F_r\left(T^*\right)\right] = C. \tag{5}$$

If the enrollment capacity constraint is not binding, i.e., if $\mu = 0$, it follows from (4) that $T_w^* = T_b^* = T^* = 0$. That is, if the capacity constraint for the elite university is not binding, it will admit all the students that will benefit from attending the elite university, but it will not admit anyone with a negative treatment effect.

2.1.2 The Second Benchmark: Asymmetric Information and No Diversity Concerns

Now consider the case where the elite university knows about a student's treatment effect T_r, but the student does not. The interpretation of the assumption that the elite university may know more about a student's treatment effect is that it has private information about the match quality between the student and the learning environment in the elite university. The key difference from the previous case is that here the student's matriculation constraint is

$$E\left[T_r | T_r \geq T_r^*\right] \geq 0. \tag{6}$$

That is, when T_r is known only to the elite university, a student upon admission only knows that her treatment effect is higher than the admission threshold adopted by the elite university.

The elite university's problem in this case can be written as:

$$\max_{\{T_r^*\}} \sum_{r\in\{w,b\}} N_r \int_{T_r^*}^{\infty} T_r f_r\left(T_r\right) dT_r \tag{7}$$

$$s.t. \quad \sum_{r\in\{w,b\}} N_r \left[1 - F_r\left(T_r^*\right)\right] \leq C, \tag{8}$$

$$E\left[T_r | T_r \geq T_r^*\right] \geq 0, \text{ for } r \in \{w, b\} \tag{9}$$

Notice that in problem (7), the elite university can in principle choose $T_r^* < 0$, as long as $E\left[T_r | T_r \geq T_r^*\right] \geq 0$, all admitted students will choose to matriculate. It is clear, however, that such choices are not optimal, because admitting students with negative treatment effects always lowers the elite university's objective. As a result, in optimum $T_r^* \geq 0$ must hold for Problem (7). Thus the solution to Problem (7) is identical to that to Problem (1).

Proposition 1 *When the elite university does not have diversity concerns, regardless of whether it has private information about students' treatment effect, its optimal admission policy is given by:*

$$T_w^* = T_b^* = \max\left\{0, T^*\right\}$$

where T^* solves (5). If $T_w^* = T_b^* > 0$, then the elite university's capacity will be full.

Notice that under the admission policy characterized in Proposition 1, the elite university is following the same admission standard for black and white students, but the racial composition of its matriculated student body may be very different from the overall composition of the applicants because $F_w(\cdot) \neq F_b(\cdot)$. If the elite university has diversity concerns, it has to modify the equal standard admission policy. We analyze these cases below.

2.1.3 The Case of Symmetric Information and Diversity Concerns

Now we suppose that the elite university has diversity concerns. First consider the case that the students know their treatment effects from attending the elite university. In this symmetric information case, no students with a negative treatment effect will matriculate in the elite university, even if they are admitted. Thus the matriculation constraint for the students must be

$$T_r^* \geq 0 \text{ for } r \in \{w, b\}.$$

The elite university's problem with diversity concerns can be written as:

$$\max_{\{T_r^* \geq 0\}} \sum_{r \in \{w,b\}} N_r \int_{T_r^*}^{\infty} T_r f_r(T_r)\, dT_r \tag{10}$$

$$s.t. \quad \sum_{r \in \{w,b\}} N_r \left[1 - F_r(T_r^*)\right] \leq C,$$

$$\frac{N_b \left[1 - F_b(T_b^*)\right]}{N_w \left[1 - F_w(T_w^*)\right]} \geq \frac{\lambda}{1 - \lambda}, \tag{11}$$

$$T_r^* \geq 0 \text{ for } r \in \{w, b\}, \tag{12}$$

where $\lambda \geq 0$ in the diversity constraint (11) stipulates that the proportion of blacks among those who matriculate in the elite university is no less than λ. We index the solutions to the above problem by $T_r^*(\lambda)$. Thus, the solution characterized in Proposition 1 is a special case, i.e., $T_r^* = T_r^*(0)$.

The solution to Problem (10) can be characterized using the standard methods, but it is useful to discuss how its solution is affected by λ. To start off, note that if the solutions $T_r^*, r \in \{w, b\}$, as characterized in Proposition 1 for the relaxed problem (1) satisfies the additional diversity constraint (11), T_r^* must also solve Problem (10). The interesting case, then, is when T_r^* characterized in Proposition 1 violates the diversity constraint (11). In that scenario, the elite university's possible responses are to decrease the admission threshold for blacks, and/or to increase the admission threshold for whites. If $T_r^* > 0$ in Proposition 1, then we know the capacity constraint binds; thus any decrease in black's admission standard must be accompanied by an increase in white's admission standard. However, under symmetric information, the students' rational enrollment constraint (12) ensures that no students with negative treatment effect will choose to enroll in the elite university, even when they are offered admission as a result of affirmative action.

Consider the interesting case where the elite university's capacity constraint binds in problem (1), and let us examine how the total treatments for blacks and for the whites are affected by the degree of diversity concern as measured by λ. For convenience, let us denote by λ_1 as the black student proportion achieved by $T_r^*(0) > 0$ for problem (1), i.e.,

$$\lambda_1 = \frac{N_b\left[1 - F_b\left(T_b^*(0)\right)\right]}{C};$$

also, denote by λ_2 as the black student proportion achieved when the admission standard for black students is set to be zero, i.e.,

$$\lambda_2 = \frac{N_b\left[1 - F_b(0)\right]}{C}.$$

Apparently, if $\lambda \leq \lambda_1$, the solution to problem (10) is the same as that for problem (1); when $\lambda \in [\lambda_1, \lambda_2]$, the solutions to problem (10) are implicitly characterized by:

$$\begin{aligned} N_b\left[1 - F_b\left(T_b^*(\lambda)\right)\right] &= \lambda C \\ N_w\left[1 - F_w\left(T_w^*(\lambda)\right)\right] &= (1-\lambda) C. \end{aligned}$$

That is, $T_b^*(\lambda)$ and $T_w^*(\lambda)$ will be chosen to satisfy exactly the capacity and the diversity constraints. When $\lambda > \lambda_2$, however, the optimal solution is to set $T_b^*(\lambda) = 0$, to choose $T_w^*(\lambda)$ to meet the diversity constraint, and leave the capacity constraint slack. That is, $T_w^*(\lambda)$ is chosen so that

$$\frac{N_b\left[1 - F_b(0)\right]}{N_b\left[1 - F_b(0)\right] + N_w\left[1 - F_w\left(T_w^*(\lambda)\right)\right]} = \lambda$$

That is

$$N_w\left[1 - F_w\left(T_w^*(\lambda)\right)\right] = \frac{1-\lambda}{\lambda}\lambda_2 C.$$

Under this admission policy, the total enrollment is given by

$$\frac{\lambda_2 C}{\lambda},$$

which is less than the allowable capacity C.

Define the total treatment effect for group r as:

$$\Phi_r(\lambda) = \int_{T_r^*(\lambda)} T_r f_r(T_r)\, dT_r.$$

Given the above discussion, we know that $\Phi_r(\lambda)$ can be depicted as in Figure 1.

Proposition 2 *When there is symmetric information about students' treatment effects, the optimal admission policy of the elite university with diversity concerns must have non-negative admission standards; and the total treatment effect of black students is non-decreasing in the degree of diversity concern as measured by λ.*

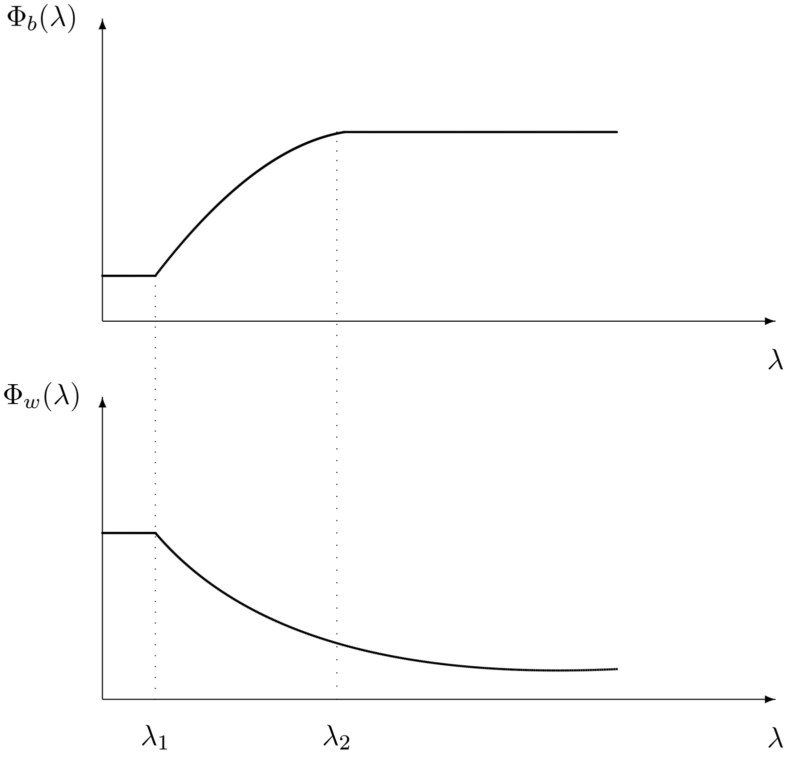

Figure 1: The Total Treatment Effects as a Function of the Diversity Concern λ: The Symmetric Information Case.

2.1.4 The Case of Asymmetric Information and Diversity Concerns

Now we consider the case where the elite university has private information about the treatment of the students. Now the elite university's optimization problem becomes:

$$\max_{\{T_r^*\}} \sum_{r\in\{w,b\}} N_r \int_{T_r^*}^{\infty} T_r f_r\left(T_r\right) dT_r \tag{13}$$

$$s.t. \quad \sum_{r\in\{w,b\}} N_r \left[1-F_r\left(T_r^*\right)\right] \leq C,$$

$$\frac{N_b\left[1-F_b\left(T_b^*\right)\right]}{N_w\left[1-F_w\left(T_w^*\right)\right]} \geq \frac{\lambda}{1-\lambda}, \tag{14}$$

$$\mathrm{E}\left[T_r|T_r \geq T_r^*\right] \geq 0 \text{ for } r \in \{w,b\}, \tag{15}$$

where as in problem (10) $\lambda \geq 0$ measures the degree of the elite university's diversity concern. Note that the only difference between problem (13) from the previous problem (10) with symmetric information lies in the difference between the student enrollment constraints (12) and (15). Under asymmetric information, the elite university can potentially attract students with negative treatment effects to enroll as long as the expected treatment effect is positive.

To characterize the solution to problem (13), it is useful to denote $\hat{T}_b < 0$ as defined by

$$\mathrm{E}\left[T_b|T_b \geq \hat{T}_b\right] = 0.$$

Furthermore, let

$$\lambda_3 = \frac{N_b\left[1-F_b\left(\hat{T}_b\right)\right]}{C};$$

that is, λ_3 is the maximal fraction of black students that can be achieved by the elite university under asymmetric information and black students' rational enrollment decisions. Note also that by definition, the total treatment effect for the blacks at λ_3 is exactly zero:

$$\Phi_b\left(\lambda_3\right) = 0.$$

Again consider the interesting case where the elite university's capacity constraint binds. The solution to problem (13) is again very simple. If the diversity concern λ is less than λ_1, the elite university does not need to modify its admission standards; if $\lambda \in (\lambda_1, \lambda_3)$, the elite university would have to lower the admission threshold for the blacks, and as a result of the capacity constraint, to increase the admission threshold for the whites correspondingly. The admission thresholds $T_r^*(\lambda)$ are again implicitly defined by

$$\begin{aligned} N_b\left[1-F_b\left(T_b^*\left(\lambda\right)\right)\right] &= \lambda C \\ N_w\left[1-F_w\left(T_w^*\left(\lambda\right)\right)\right] &= \left(1-\lambda\right) C. \end{aligned}$$

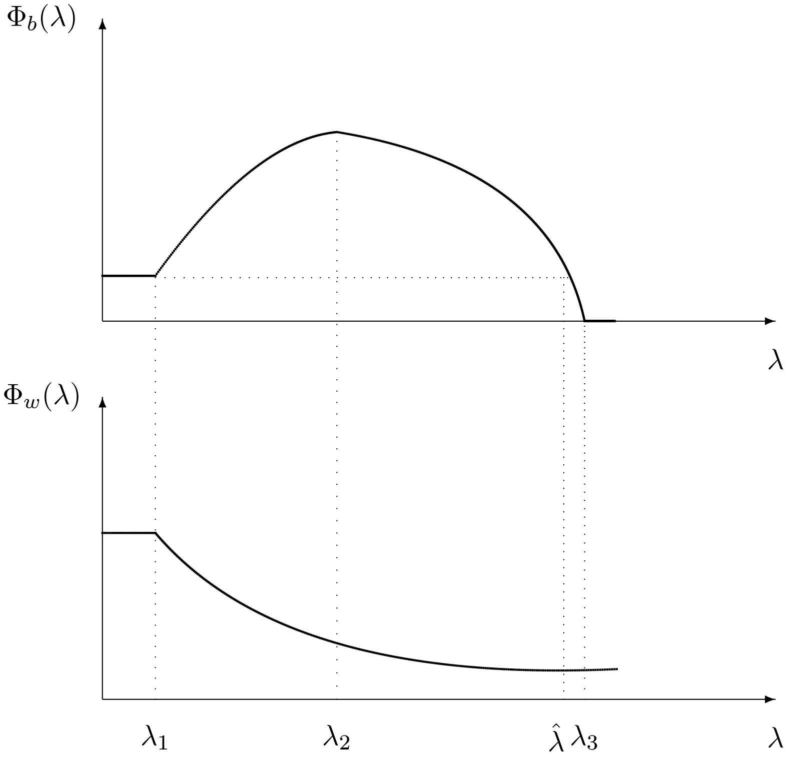

Figure 2: The Total Treatment Effects as a Function of the Diversity Concern λ: The Asymmetric Information Case.

When $\lambda > \lambda_3$, the elite university can no longer increase black enrollment by lowering the admission standard because of the binding enrollment constraint (15). Thus the only way it can satisfy the diversity constraint is to admit fewer white students. As a result, when $\lambda > \lambda_3$, the elite university's total enrollment will be

$$\frac{\lambda_3 C}{\lambda},$$

which is less than the allowable capacity C. The effect of the diversity concern λ on the total treatments of black and white students in this case is depicted in Figure 2. Note that the key difference between Figure 1 (the symmetric information case) and Figure 2 (the asymmetric information case) is that in the asymmetric information case, increases in λ may lead a decrease of the black total treatment effect relative to the case with no diversity concerns $(\lambda = 0)$. In fact, the total black treatment effects are smaller than those with no diversity concerns for $\lambda > \hat{\lambda}$ where $\Phi_b\left(\hat{\lambda}\right) = \Phi_b(0)$.

The following proposition summarizes the key results from this section:

Proposition 3 *In the asymmetric information case, the elite university's admission threshold for the black students, $T_b^*(\lambda)$, is strictly decreasing in the extent of the diversity concern λ as long as $\lambda \leq \lambda_3$. However, the total treatment effect for the blacks, $\Phi_b(\lambda)$, is not monotonic in λ. In particular, when $\lambda > \hat{\lambda}$, $\Phi_b(\lambda) < \Phi_b(0)$.*

2.2 Mismatch and Asymmetric Information

Now we are ready to present our main conclusion from the analysis so far. First, let us provide several notions of "mismatch" as a result of affirmative action admission policies by the elite university.

Definition 1 *We say that affirmative action admission policy by the elite university leads to a **local mismatch effect** for the blacks if some black students with negative treatment effect are admitted and enrolled, that is, if $T_b^*(\lambda) < 0$.*

Definition 2 *We say that affirmative action admission policy by the elite university leads to a **global mismatch effect** for the blacks if the average black students are made worse off, i.e.,*

$$\int_{T_b^*(\lambda)} T_b dF_b(T_b) < \int_{T_b^*(0)} T_b dF_b(T_b). \tag{16}$$

Equivalently, (16) can be written as

$$\mathrm{E}\left[T_b | T_b \geq T_b^*(\lambda)\right]\left[1 - F_b\left(T_b^*(\lambda)\right)\right] < \mathrm{E}\left[T_b | T_b \geq T_b^*(0)\right]\left[1 - F_b\left(T_b^*(0)\right)\right].$$

Note from Proposition 1, $T_b^*(0) = T_b^* \geq 0$ regardless of whether the elite university has asymmetric information about the students' treatment effects. Together with the fact that $T_b^*(\lambda)$ is weakly decreasing in λ, we can conclude that a global mismatch is possible only if $T_b^*(\lambda)$ is sufficiently negative. Thus global mismatch must imply local mismatch.

Because both the local and global notions of mismatch require that the admission thresholds for blacks, $T_b^*(\lambda)$, to be sufficiently negative, and students with negative treatment effect will choose to attend the elite university only when they are not fully knowledgeable about their treatment effect, we conclude that a *necessary* condition for mismatch to occur is that the elite university has private information regarding the students' treatment effect. Combining the results from Propositions 2 and 3, we have:

Proposition 4 *A* ***necessary condition*** *for either local or global mismatch to result from affirmative action admission policy is that the elite university has private information about the students' treatment effect.*

A corollary is that if we do not find evidence that elite universities have private information about students' treatment effect, then there is no scope of mismatch when we interpret mismatch from utility-based viewpoint.

2.3 Robustness of the Main Results

So far the elite university's diversity concern is modelled as an absolute constraint (see 11 and 14). Now we show that our main result connecting mismatch to private information does not depend on such formulations of the diversity concern. The key driver of the result is the rational student enrollment constraint.

To see this, suppose that the diversity concern is incorporated into the elite university's objective function, instead of its constraints. Suppose that the objective function of the elite university is

$$\sum_{r\in\{w,b\}} N_r \int_{T_r^*}^{\infty} T_r f_r(T_r)\, dT_r - \alpha \left\{(1-\lambda) N_b \left[1 - F_b(T_b^*)\right] - \lambda N_w \left[1 - F_w(T_w^*)\right]\right\}^2$$

where $\alpha > 0$; and the term multiplied by α captures the payoff loss if the enrolled black/white student ratio is far from the desired ratio of $\lambda/(1-\lambda)$.

If there is symmetric information about the treatment effects, the elite university's problem is

$$\begin{aligned}
&\max_{\{T_r^*\}} \sum_{r\in\{w,b\}} N_r \int_{T_r^*}^{\infty} T_r f_r(T_r)\, dT_r - \alpha \left\{(1-\lambda) N_b \left[1 - F_b(T_b^*)\right] - \lambda N_w \left[1 - F_w(T_w^*)\right]\right\}^2 \\
s.t. \quad & \sum_{r\in\{w,b\}} N_r \left[1 - F_r(T_r^*)\right] \leq C \\
& T_r^* \geq 0, \text{ for } r \in \{w,b\}.
\end{aligned}$$

As in problem 10, $T_b^* \geq 0$ ensures that there is no scope for mismatch to happen in either local or global sense.

If there is asymmetric information about the treatment effects, the elite university's problem is

$$\begin{aligned}
&\max_{\{T_r^*\}} \sum_{r\in\{w,b\}} N_r \int_{T_r^*}^{\infty} T_r f_r(T_r)\, dT_r - \alpha \left\{(1-\lambda) N_b \left[1 - F_b\left(T_b^*\right)\right] - \lambda N_w \left[1 - F_w\left(T_w^*\right)\right]\right\}^2 \\
s.t. \quad &\sum_{r\in\{w,b\}} N_r \left[1 - F_r\left(T_r^*\right)\right] \leq C \\
&E\left[T_r | T_r \geq T_r^*\right] \geq 0, \text{ for } r \in \{w, b\}.
\end{aligned}$$

It is easy to see that when λ is sufficiently large, it is possible that the optimal solution will have $T_b^* < 0$.

3 An Empirical Method for the Identification of Private Information

In Section 2, we argued that once we take into account the students' rational enrollment decisions, a *necessary* condition for either local mismatch or global mismatch to arise as a result of affirmative action admission policies by the elite university is that the elite university has private information about the students' treatment effect. In this section, we propose tests for private information by the elite university. If our test rejects the presence of private information by the elite university, then we can conclude that mismatch does not arise as a result of affirmative action admission policies; however, if we detect private information, it is *not sufficient* to establish that mismatch occurred. Our findings, however, will always have policy implications even when we can not conclude whether mismatch occurs. We discuss these policy implications in the conclusion.

There is a large existing economics literature that tests for asymmetric information particularly for adverse selection in the empirical analysis of a variety of insurance markets.[6] Most of these papers test whether the data supports a positive association between insurance coverage and *ex post* risk occurrence, a robust prediction of the classical models of insurance market developed by Arrow (1963), Pauly (1974), Rothschild and Stiglitz (1976) and Wilson (1977).[7]

Our setting substantially differs from the insurance market setting studied in the existing literature. The empirical insurance literature assumes that private information is possessed by one-side

[6] The rapidly growing literature includes Cawley and Philipson (1999) for life insurance market, Chiappori and Salanie (2000) for auto insurance market, Cardon and Hendel (2001) for health insurance market, Finkelstein and Poterba (2004) for annuity market, Finkelstein and McGarry (2006) for long-term care insurance market and Fang, Keane and Silverman (2008) for Medigap insurance market.

[7] See Chiappori et. al. (2006) for a general derivation of the positive association property.

of the market, the potential insured, and it is manifested through their insurance purchase and their ex post risk occurrence. In our setting, there presumably is private information about the treatment effect by both the student and the university. Moreover, the empirical insurance literature typically assumes either to have access to observations for individuals with and without insurance and their risk realizations, or to have access to observations for individuals with different amount of coverage and their risk realizations. In particular, the risk realization may be related to insurance coverage due to moral hazard, but will be unrelated to which insurance company provides the coverage. In our setting, if a student does not attend the elite university, we will not observe the student's outcome had he attended it; or if the student attends the elite university, we will not observe the student's outcome had he not attended. For these reasons we will describe below an empirical strategy to identify private information in our setting.

3.1 Available Data and Assumptions

Suppose that we have data about the **observed student outcome** Z. Z could be student's GPA, or post-graduate income etc. Conceptually, we assume that Z is a linear function of X_U, X_S and X_C where X_U denotes the unobserved university's private information about student performance, X_S denotes the unobserved student's private information and X_C denotes the information that is common to both students and the university but unobserved by the researcher. Of course, we can also include a set of variables Y that are common information to the university and the students and are observed by researchers, such as observed family and high school characteristics; we will ignore Y for the discussion here for simplicity.

Specifically, suppose that

$$Z = \alpha_C X_C + \alpha_U X_U + \alpha_S X_S + \varepsilon, \tag{17}$$

where ε is noise. By construction, and thus without loss of generality we assume that X_C, X_U, X_S and ε are independent.

We assume that we also have access to **two additional variables**: a variable, denoted by W_U, that measures the selective university's assessment about the student's treatment effect given its private knowledge about the match between the student and the university X_U, as well as the common information X_C; and another variable denoted by W_C that measures the student's own performance expectation in the selective university given the common information X_C and her own private information X_S. We assume that (W_U, W_S) are related to X_C, X_U and X_S as follows:

$$\begin{aligned} W_U &= X_C + X_U, && (18) \\ W_S &= X_C + X_S. && (19) \end{aligned}$$

To summarize, we assume that we observe a data set consisting $\{W_U, W_S, Z\}$ and assume that there exists independent variable variables X_C, X_U, X_S and ε such that $\{W_U, W_S, Z\}$ are generated by (17)-(19).

The question we are interested in is, how do we estimate the coefficients α_C, α_U and α_S, and/or decompose the importance of common information X_C, student private information X_S, university private information X_U and noise ε in explaining the variation of Z in the data?

3.2 An Empirical Strategy

We propose an empirical strategy that consists of the following steps:

1. Invoking Kotlarski's (1967) theorem, we separately recover the marginal distributions of X_C, X_U and X_S from the observed joint distribution of (W_U, W_S);
2. We draw random samples of $\left\{X_C^i, X_U^i, X_S^i\right\}$ from the marginal distributions of X_C, X_U and X_S recovered in step 1;
3. We obtain samples of $\left\{W_U^i, W_S^i\right\}$ from the random samples of $\left\{X_C^i, X_U^i, X_S^i\right\}$ generated in step two, and then draw sample of Z^i conditional on $\left\{W_U^i, W_S^i\right\}$ from the observed conditional distribution $G\left(Z|W_U, W_S\right)$;
4. We run regressions of Z on X_C, X_U, X_S using the sample $\left\{Z^i, X_C^i, X_U^i, X_S^i\right\}$ simulated above to estimate α_C, α_U and α_S, and to do variance decomposition.

Now we provide more details about the above empirical strategy. The key is the first step. The key mathematical result we use is the Kotlarski's theorem:

Theorem 1 ***(Kotlarski's Theorem)*** *Let X_C, X_U and X_S be three independent real-valued random variables. Suppose W_U and W_S are generated as in (18) and (19). Then the joint distribution of (W_U, W_S) determines the marginal distribution of X_C, X_U, X_S up to a change of the location as long as the characteristic function of (W_U, W_S) does not vanish (i.e., it does not turn into zero on any non-empty interval of the real line).*

This well-known theorem is first proved in Kotlarski (1967) and the proof can also be found in Rao (1992, pp 7-8).[8] The proof of the theorem also suggests how the marginal distributions for

[8] Kotlarski theorem has been widely used in measurement error models in econometrics (e.g., Li and Vuong 1998). It has been applied elsewhere in economics, e.g. Krasnokutskaya (2004) used in the context of identifying and estimating auction models with unobserved auction heterogeneity, and Cunha, Heckman and Navarro (2005) used it to distinguish uncertainty from heterogeneity in their analysis of life-cycle earnings.

X_C, X_S and X_U can be constructed. Let

$$\Psi\left(t_1, t_2\right) = \mathrm{E}\exp\left(it_1 W_U + it_2 W_S\right) \tag{20}$$

denote the characteristics function for the observed joint random vector $\left(W_U, W_S\right)$, and let

$$\begin{aligned} \Psi_1\left(t_1, t_2\right) &\equiv \frac{\partial \Psi\left(t_1, t_2\right)}{\partial t_1} \\ &= \mathrm{E}\left[iW_U \exp\left(it_1 W_U + it_2 W_S\right)\right] \end{aligned} \tag{21}$$

denote the derivative of $\Psi\left(\cdot,\cdot\right)$ with respect to its first argument. Then Kotlarski theorem shows that the characteristic function for random variables X_C, X_U, X_C are respectively given by

$$\begin{aligned} \Psi_{X_C}\left(t\right) &= \log\left(\int_0^t \frac{\Psi_1\left(0, t_2\right)}{\Psi\left(0, t_2\right)} dt_2\right), \\ \Psi_{X_U}\left(t\right) &= \frac{\Psi\left(t, 0\right)}{\Psi_{X_C}\left(t\right)}, \\ \Psi_{X_S}\left(t\right) &= \frac{\Psi\left(0, t\right)}{\Psi_{X_C}\left(t\right)}. \end{aligned}$$

Finally the characteristic functions of these three random variables uniquely determines the probability density function via an inversion formula. Let f_{X_C}, f_{X_U}, and f_{X_S} respectively denote the marginal probability density function for random variables X_C, X_U and X_S. We have, following the inversion formula described in Horowitz (1998, pp. 104)

$$f_{X_K}\left(x_K\right) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \exp\left(-itx_K\right) \Psi_{X_K}\left(t\right) dt \text{ for } K \in \left\{C, U, S\right\}.$$

Once we have the marginal distributions for X_K for $K \in \left\{C, U, S\right\}$, the remaining steps 2-4 described above are rather straightforward. Now we describe the somewhat standard estimation procedure to carry out step 1.[9] The key is to estimate $\Psi\left(\cdot,\cdot\right)$ and $\Psi_1\left(\cdot,,\cdot\right)$ by their sample analogs: given a sample $\left\{W_U^j, W_S^j\right\}_{j=1}^n$,

$$\begin{aligned} \Psi\,\widehat{\left(t_1, t_2\right)} &= \frac{1}{n}\sum_{j=1}^n \exp\left(it_1 W_U^j + it_2 W_S^j\right) \\ \Psi_1\,\widehat{\left(t_1, t_2\right)} &= \frac{1}{n}\sum_{j=1}^n iW_U^j \exp\left(it_1 W_U^j + it_2 W_S^j\right). \end{aligned}$$

The characteristic functions $\Psi_{X_K}\left(t\right)$ for $K \in \left\{C, U, S\right\}$ can in turn be estimated by replacing $\Psi\left(\cdot,\cdot\right)$ and $\Psi_1\left(\cdot,,\cdot\right)$ by their estimates above.

[9] See Krasnokutskaya (2004) for similar estimation procedure. Horowitz (1998, Chapter 4) describes some useful suggestions for issues related to smoothing.

3.3 Some Remarks

We have assumed in equation (17) that the student outcome Z is a linear function of X_C, X_U, X_S. This is for simplicity only. With the pseudo data sets we simulated in Step 3, we can also estimate Z as a nonlinear function of these variables, or even nonparametrically estimate their relations.

It is also worth noting in specification (18) and (19), we interpret X_U and X_S are respectively the true private information for the university and the student, and assume away noise in the measurement variables W_U and W_S. If instead the variables we extract in step 1 contain the true private information of the university and students contaminated by noise, then we will have, in step 4, a mismeasured independent variables in the regressions. This may bias our coefficient estimates for α_U and α_S downward, but when we do variance decomposition for Z, we should still be able to recover the importance of the true private information of the university and the student in explaining the variance of the outcome variable Z.

3.4 Monte Carlo Results

[**To be completed**] Here we will add results from Monte Carlo for the estimation algorithm proposed above.

4 The Campus Life and Learning (CLL) Project Data

In this section, we describe a data set from the Campus Life and Learning Project (CLL) at Duke that fits into the data requirement to implement the empirical strategy we described above.[10] CLL is a multi-year prospective panel study of consecutive cohorts of students enrolled at Duke University in 2001 and 2002 (graduating classes of 2005 and 2006).[11] The target population of the CLL project included all undergraduate students in Duke's Trinity College of Arts & Sciences and Pratt School of Engineering. Using the students' self-reported racial ethnic group from their Duke Admissions application form, the sampling design randomly selected about 356 and 246 white students from the 2001 and 2002 cohorts respectively, all black and Latino students, about two thirds of Asian students and about one third of Bi-Multiracial students in each cohort. The final design across both cohorts contains a total of 1536 students, including 602 white, 290 Asian, 340 black, 237 Latino and 67 Bi-Multiracial students.

[10] A description of the CLL Project and its survey instruments can be found at `http://www.soc.duke.edu/undergraduate/cll/`, where one can also find the reports by Bryant et. al. (2006, 2007).

[11] Duke is among the most selective national universities with about 6,000 undergraduate students. Duke's acceptance rate for its regular applications is typically less than 20 percent.

			By Race							
Variable	Full Sample ($N = 940$)		White ($N = 419$)		Black ($N = 174$)		Asian ($N = 178$)		Latino ($N = 169$)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Duke Admission Office Evaluations:										
Achievement	4.25	0.85	4.34	0.87	3.75	0.80	4.67	0.58	4.13	0.81
Curriculum	4.70	0.55	4.71	0.56	4.46	0.62	4.91	0.37	4.72	0.50
Essay	3.45	0.58	3.52	0.61	3.26	0.48	3.58	0.66	3.31	0.52
Personal Qualities	3.47	0.60	3.57	0.64	3.34	0.51	3.52	0.63	3.30	0.51
Recommendations	3.83	0.66	3.97	0.68	3.55	0.60	4.06	0.57	3.55	0.57
Test Scores	3.31	1.36	3.69	1.18	2.09	1.04	4.10	1.13	2.79	1.23
Student's Expected First Year GPA	3.53	0.32	3.51	0.31	3.44	0.34	3.67	0.28	3.53	0.31
Actual First Year Cum. GPA	3.22	0.49	3.33	0.46	2.90	0.48	3.40	0.42	3.13	0.46
Graduate Early or On Time	0.89	0.31	0.90	0.30	0.88	0.33	0.92	0.27	0.87	0.34
Private High School	0.31	0.46	0.31	0.46	0.25	0.44	0.25	0.44	0.39	0.49
SAT (Math + Verbal)	1388.12	115.19	1416.63	99.68	1280.96	92.51	1463.66	90.24	1349.03	101.99
Family Income Less than $49,999	0.18	0.39	0.10	0.30	0.32	0.47	0.19	0.39	0.22	0.42
Family Income $50K to $99,999	0.23	0.42	0.19	0.40	0.30	0.46	0.24	0.43	0.23	0.42
Family Income Above $100K	0.59	0.49	0.71	0.46	0.37	0.49	0.57	0.50	0.54	0.50
Father Education BA or Higher	0.84	0.37	0.91	0.29	0.63	0.48	0.90	0.30	0.78	0.42
Mother Education BA or Higher	0.74	0.44	0.82	0.39	0.61	0.49	0.74	0.44	0.70	0.46

Table 1: Summary Statistics of Key Variables. NOTE: These summary statistics are unweighted.

Each cohort was surveyed via mail in the summer before initial enrollment at Duke, in which they were also asked to sign an informed consent document, as well as given option of providing confidential access to their student information records at Duke. About 78 percent of sample members ($n = 1185$) completed the pre-college mail questionnaire; and 98.2 percent of the respondents provided signed release to their institutional records for the study. In the spring semester of the first, second and fourth college year, each cohort was again surveyed by mail.[12] However, response rates declined in the years following enrollment: the response rates were respectively 71, 65 and 59 percent in the first, second and fourth college years.

The pre-college survey provides detailed measurement of the students' social and family background, prior school experiences, social networks, and importantly, their expectations for college. The in-college surveys contain data on social networks, performance attributions, choice of major, residential and social life, perception of campus climate and plans for the future. For those consented access to their institutional records, we have information about their grades, graduation outcomes, all relevant information in their Duke application including test scores (i.e. SAT, ACT), Duke Admission Officers' rankings about their application based on high school curriculum, reader rating scores, high school extracurricular activities, and financial aid and support.

The empirical strategy we proposed in Section 3 above requires that we have data about some university assessment about the student's treatment effect (W_U), and a student assessment about their own treatment in the university (W_S), as well as an outcome measure of the student in the university (Z). We proxy W_U with Duke Admission Officers' rankings of the applicant, and proxy W_S by the students' answer to the question in the pre-college survey:

> "What do you realistically expect will be your cumulative GPA at Duke after your first year?"

And finally, we use the students' cumulative GPA at the end of first year to proxy Z.[13] We will include standard controls for family and high school background such as parents' education, type of high school (i.e. private or public), family income, etc.

Table 1 contains summary statistics for the key variables in our CLL data set. We will use six Duke Admission Office Evaluations on achievement (DukeEv-A), curriculum (DukeEv-C), essay (DukeEv-E), personal qualities (DukeEv-P), recommendations (DukeEv-R) and test scores

[12] The survey was not conducted in the third year as many Duke students are studying abroad during that year.

[13] We have the students' cumulative GPA at all years. We choose to use the first year cumulative GPA because the courses that the students had much less discretion to choose "easy" classes in their freshman year.

(DukeEv-T) as W_U. All of these variables are scaled between 1 to 5.[14] Simple calculations of the means show that there is substantial amount of variations among the entering students with difference races: Asians and Whites tend to have higher evaluations by Duke Admission Officers in all six categories than black and Latino students. On the variable that we use to proxy for W_S, the student's first year expected GPA (ExpGPA), Table 1 shows that, on average, black and white students have quite similar projections about their expected GPA during their first year in college (i.e. 3.51 for whites and 3.44 for blacks).[15]

However, Table 1 shows that there is a significant racial difference in the actual first year cumulative GPAs (CumGPA). The actual GPA for blacks is on average 2.90, in contrast to that for whites (3.33) and for Asians (3.40). In fact a t-test rejects the null hypothesis of equal means. Notice that, for all races, the students' actual first year GPAs are on average lower than their expected GPAs. This suggests that all students have over-optimistic expectations; however, this optimism bias is much stronger for black (0.54) and Latino (0.4) students than for white (0.18) and Asian (0.27) students. Again, a t-test rejects the null hypothesis of equality of means.

Of course, part of the actual GPA differences across races are predicted by observable differences across races in their entering credentials. For example, Table 1 shows, for example, Asians and whites have substantially higher (more than one standard deviations) SAT scores than Latino and black students. Average family income for Black students tend to be lower than Asians and Latinos, which in turn are lower than the whites. The parents of white students tend to have higher educational attainment than blacks. It is interesting to note that, even though whites are more likely to attend private high school, Asian students are equally likely to have attended a public high school as black students.

The key question is then, why do the black and Latino students suffer a worse bias in their expectation about their academic performance at Duke? Does Duke Admission Office's evaluation of their application contain valuable information that would have been useful in helping these students forming more realistic expectations? If the black and Latino students were able to form more realistic expectations about their academic performance at Duke, would they have reconsidered their decisions to enroll at Duke? These are the key empirical questions related to the mismatch hypothesis.

[14] Because it is not clear how a scale measure of 1 compares with a measure of 2, we will use polychoric transformations of these scale measures in our empirical section.

[15] A t-test *cannot* reject the null hypothesis of equal means.

5 Results (Preliminary)

In this section we present some preliminary results using two empirical strategies that differ from the nonparametric methods we proposed in Section 3.[16]

5.1 Strategy 1: Factor Analysis

In our first strategy, we use factor analysis to decompose the six Duke Admission Office variables (DukeEv-A, DukeEv-C, DukeEv-E, DukeEv-P, DukeEv-R and DukeEv-T) and the student's expected first year GPA (ExpGPA) into a small number of common factors and unique factors.[17] Specifically, we attempt to extract factors F_κ, $\kappa = 1, ..., K$, with K to be determined empirically, such that, for each J in the set {A,C,E,P,R,T} there exists some coefficients $\beta_{i,\mathrm{J}}$ such that

$$\begin{aligned} \text{DukeEv-J}^i &= \sum_{\kappa=1}^{K} \beta_{\kappa,\mathrm{J}} F_\kappa^i + u_\mathrm{J}^i, \\ \text{ExpGPA}^i &= \sum_{\kappa=1}^{K} \beta_{S,\kappa} F_\kappa^i + u_S^i \end{aligned}$$

for each student i. Now the factors F_κ will be interpreted as common information, and the uniqueness factors u_J will be interpreted as the private information of the university and the uniqueness factor u_S is considered as the private information of the student.

Factor analysis of the six Duke evaluation variables DukeEv-J, for J $\in$ {A,C,E,P,R,T} and ExpGPA results up to three common factors. In Table 2, we report the coefficient estimates from regressions of CumGPA on the common factors extracted from factor analysis, the unique factors, as well as a list of the students' observed personal, family and high school characteristics (see notes for Table 2 for these additional controls). The six unique factors for DukeEv-J variables from the factor analysis are combined into a single measure of Duke private information; and the unique factor from ExpGPA is used to measure the student's private information. The three columns in Table 2 differ in the number of retained common factors. Because the unique factors differ depending on how many common factors are retained, we report the standard deviations for Duke and student private information separately for the three different regression specifications.

The main messages from Table 2 are as follows. First, even after controlling for the list of personal, family and high school characteristics, the common factors we extracted from the factor

[16] We are in the process of evaluting the nonparametric methods using Monte Carlo, followed by actual implementation using the CLL data.

[17] See Kim and Mueller (1994) for an excellent theoretical introduction and Cureton and D'Agostino (1983) for practical issues of factor analysis.

	One Factor	Two Factors	Three Factors
Variables	$(K = 1)$	$(K = 2)$	$(K = 3)$
	(1)	(2)	(3)
Factor 1 (F_1) [s.d.=0.72]	0.296^{***}	0.250^{***}	0.258^{***}
	(0.033)	(0.034)	(0.037)
Factor 2 (F_2) [s.d.=0.68]		0.119^{***}	0.123^{***}
		(0.024)	(0.024)
Factor 3 (F_3) [s.d.=0.28]			-0.135^{**}
			(0.056)
Duke Priv. Info. (One-Factor Model)	1.020^{***}		
[s.d.=0.104]	(0.171)		
Duke Priv. Info. (Two-Factor Model)		0.960^{***}	
[s.d.=0.076]		(0.241)	
Duke Priv. Info. (Three-Factor Model)			1.020^{**}
[s.d.=0.065]			(0.412)
Student Priv. Info. (One-Factor Model)	0.041		
[s.d.=0.263]	(0.067)		
Student Priv. Info. (Two-Factor Model)		0.035	
[s.d.=0.258]		(0.069)	
Student Priv. Info. (Three-Factor Model)			0.046
[s.d.=0.254]			(0.094)
Constant	3.196^{***}	3.197^{***}	3.197^{***}
	(0.062)	(0.062)	(0.062)
Adjusted R^2	0.3201	0.300	0.299
Obs.	802	802	802

Table 2: Does Duke Possess Private Information that Can Be Used to Predict the Students' Actual First Year Cumulative GPA?

Notes: The dependent variable is CumGPA. Standard errors are reported in parenthesis. All regressions are OLS, controlling for students' observed personal, family and high school characteristics, including gender, race, father's education, family income, private school and demeaned SAT score.

The "s.d." in square bracket denotes the standard deviation of the variable.

*, **, *** respectively represents significance at 10%, 5% and 1%.

	Model		
Variables	One Factor	Two Factors	Three Factors
Common Info. (obs. & unobs.)	28.55%	30.54%	31.24%
Duke Priv. Info.	4.96%	2.38%	2.05%
Student Priv. Info	1.01%	0.51%	0.07%
Noise	65.48%	66.57%	60.01%

Table 3: Variance Decomposition for Student's First Year Cumulative GPA (CumGPA): Factor Analysis.

analysis remain significant predictors of the students' actual GPA at the end of their first year. This suggests that Duke Admission Officers glean valuable information from the students' essays, personal experiences, recommendation letters, etc. that are not reflected in other observable student characteristics. That is, there is substantial common information between students and Duke that are not observed by researchers. Second, it also shows that our measure of Duke private information is also a statistically significant predictor of CumGPA, while the measure of student private information is not. In the one factor model (column 1), a one-s.d. increases in Duke private information predicts about 0.1 increase in the student's actual GPA at the first year.

While Table 2 shows that our measure of Duke private information is a statistically significant predictor of the students actual GPA, its importance in explaining the variations in CumGPA is not clear. Table 3 reports the variance decomposition of CumGPA. It shows that regardless of whether we retain one or two or three common factors in the factor analysis, the observed and unobserved common information jointly explain about 30% of the variance in CumGPA; Duke private information explains between 2 to 5 percent of the variance; and the student's private information explains no more than 1 percent. The fraction of CumGPA explained by the student's private information ranges between 1/3 to 1/4 of the fraction explained by Duke private information. More than 60% of the variance of CumGPA is simply noise.

The conclusion from Table 2 and 3 is that, first, Duke does possess useful private information that can help predict students' academic performance after enrolling, but Duke's private information does not seem to be able to explain a substantial amount of the students' variations in their first year GPA.

5.2 Strategy 2: An Alternative Decomposition Strategy

Here we also present the results from an alternative, maybe somewhat more naive, strategy to decompose the information into common, Duke private and student private information. The decomposition is based on three separate regressions:

$$\begin{aligned} \textsc{CumGPA}^i &= \boldsymbol{\mu}_{\mathbf{x}}\mathbf{X}^i + \sum \mu_{\textsc{j}}\textsc{DukeEv-j}^i + \mu\textsc{ExpGPA}^i + \eta_1^i, & (22) \\ \textsc{CumGPA}^i &= \boldsymbol{\delta}_{\mathbf{x}}\mathbf{X}^i + \delta\textsc{ExpGPA}^i + \eta_2^i, & (23) \\ \textsc{CumGPA}^i &= \boldsymbol{\gamma}_{\mathbf{x}}\mathbf{X}^i + \sum \gamma_{\textsc{j}}\textsc{DukeEv-j}^i + \eta_3^i, & (24) \end{aligned}$$

where **X** denotes a list of controls for the student's personal, family and high school characteristics. Note that (22) regresses the observed actual cumulative GPA on a **X**, all Duke Admission Office evaluation variables DukeEv-j, as well as the student's pre-enrollment expected GPA (ExpGPA); we interpret the residual from this regression $\hat{\eta}_1$ as a measure of the noise term in (17). In contrast, (23) excludes DukeEv-j variables from the regression and we interpret the residual $\hat{\eta}_2$ as a measure of Duke private information *plus* noise; similarly (24) excludes ExpGPA from the regression and we interpret the residual $\hat{\eta}_3$ as a measure of student private information *plus* noise. Thus in this alternative strategy, we obtain the following decomposition:

$$\begin{aligned} \text{Noise } (\varepsilon) &= \hat{\eta}_1, \\ \text{Duke Priv. Info. } (X_U) &= \hat{\eta}_2 - \hat{\eta}_1, \\ \text{Student Priv. Info. } (X_S) &= \hat{\eta}_3 - \hat{\eta}_1, \end{aligned}$$

Finally, the unobserved common information X_C can be retrieved, using (19), as:

$$\text{Unobserved Common Info. } (X_C) = \textsc{ExpGPA} - (\hat{\eta}_3 - \hat{\eta}_1).$$

Using the variables for Duke and student private information, as well as the unobserved common information, constructed as explained above, the left panel in Table 4 reports the estimated coefficients from regressing CumGPA on these variables, controlling for a list of student characteristics **X**. Consistent with the findings using factor analysis reported in Table 2, we also find that after controlling for **X**, Duke private information as measured here remains a statistically significant predictor of the students' actual GPA. In terms of magnitude, a one-standard deviation increase in Duke private information increases the predicted value of CumGPA by about 0.198. The measured student private information, as in Table 2, is neither a statistically significant nor a numerically significant predictor for CumGPA. What differs from Table 2 is that under strategy the unobserved common information is no longer statistically significant.

Regression Results			Variance Decomposition for CumGPA	
Variable	Coeff. Est.	Std. Err.	Variable	Fraction (%)
Unobs. Common Info. [s.d. = 0.305]	0.129	0.173	Common Info. (unobs. & obs.)	24.94
Duke Priv. Info. [s.d. = 0.143]	1.000^{***}	0.124	Duke Priv. Info.	8.57
Student Priv. Info. [s.d. = 0.016]	0.129	4.004	Student Priv. Info.	0.32
Constant	2.274^{***}	0.604	Noise	66.17
Adjusted R^2	0.3137			
No. Obs.	802			

Table 4: Regression Coefficients and Variance Decomposition: Alternative Strategy.

NOTES: The dependent variable is CumGPA. All regressions are OLS, controlling for students' observed personal, family and high school characteristics, including gender, race, father's education, family income, private school and demeaned SAT score.

The "s.d." in square bracket denotes the standard deviation of the variable.

*, **, *** respectively represents significance at 10%, 5% and 1%.

The right panel in Table 4 undertakes variance decomposition for CumGPA similar to what is reported in Table 3. Again, the common information, both observed and unobserved, together explains about 25% of the variation, Duke private information explains an additional 8.57% of the variance, student private information explains very little, and about 2/3 of the variation in CumGPA is just noise.

5.3 Issues of the Above Strategies

Using two strategies to decompose Duke and student private information from unobserved common information, we have obtained, overall, qualitatively similar results: Duke possesses private information that is a statistically significant predictor of the students' academic performance after enrollment, but it explains a relatively small percent of the variation in the observed cumulative GPA; students, however, do not seem to possess any private information regarding their performance that is not contained in their observable characteristics.

These strategies, however, both suffer from similar problems: there is no guarantee that the measured Duke private information, student private information and unobserved common information, are orthogonal to each other. As a result, it does not exactly fit into the assumed data generating processes we outlined in Section 3.

Given this, it should be noted at this point that the variance decompositions we reported above are obtained by sequentially adding the variables and the additional variance explained is reported

Regressions	R^2		Duke Priv. Info.	Student Priv. Info.
(25)	0.2410	Upper Bound	9.2%	0.8%
(26)	0.3330	Lower Bound	8.61%	0.21%
(27)	0.2490			
(28)	0.3351			

Table 5: Bounds on the Contributions of Duke and Student Private Information to the Variation in CumGPA.

as the contribution of the variable (except for "noise", which receives simply the unexplained share of the variance). It is important to emphasize also that the results by changing the order in which Duke private information and student private information are added in the variance decomposition only causes very little changes in any of the magnitude.

Now we report results from another approach that may provide bounds on how much of the variation in CumGPA could be explained by Duke and student private information respectively. The bounds for the contribution of Duke private information are obtained from the following series of regressions:

$$\begin{aligned}
\textsc{CumGPA}^i &= \boldsymbol{\mu}_{\mathbf{x}}\mathbf{X}^i + \eta_1^i, && (25)\\
\textsc{CumGPA}^i &= \boldsymbol{\beta}_{\mathbf{x}}\mathbf{X}^i + \sum \beta_{\textsc{j}} \textsc{DukeEv-j}^i + \eta_2^i, && (26)\\
\textsc{CumGPA}^i &= \boldsymbol{\delta}_{\mathbf{x}}\mathbf{X}^i + \delta \textsc{ExpGPA}^i + \eta_3^i, && (27)\\
\textsc{CumGPA}^i &= \boldsymbol{\gamma}_{\mathbf{x}}\mathbf{X}^i + \sum \gamma_{\textsc{j}} \textsc{DukeEv-j}^i + \gamma \textsc{ExpGPA}^i \eta_4^i, && (28)
\end{aligned}$$

The difference in R^2 between regressions (25) and (26) [respectively, between regressions (27) and (28)] tells us the upper [respectively, lower] bound on the proportion of variance in CumGPA that could be explained by Duke private information. Analogously, the difference in R^2 between regressions (25) and (27) [respectively, between regressions (26) and (28)] tells us the upper [respectively, lower] bound on the proportion of variance in CumGPA that could be explained by student private information. Table 5 shows that using this method, Duke private information accounts for somewhere between 8.61% and 9.2% of the variation in CumGPA while student private information accounts for 0.21% to 0.8%. Thus qualitatively the conclusion that Duke does possess private information that can predict the students' post-enrollment performance is robust.

We believe that the Kotlarski decomposition strategy we outlined in Section 3 will provide us with decomposition of the three pieces of information that satisfy the orthogonality condition.

6 Discussions

We have argued that for affirmative action to lead to mismatch effect in the sense that its intended beneficiary may be made worse off, a necessary condition is that the selective university has private information about the student's treatment effect. However, we can not conclude from the evidence for the selective university's possession of substantial private information that may help predict the student's outcome that there is a mismatch effect.

It is worth noting, however, even if one cannot conclusively prove the existence of mismatch effect, an evidence that a selective university possesses valuable *ex ante* information that can help predict the students' academic performance should lead to cautions for possible mismatch. Recall that in our framework, mismatch from a utilitarian sense with fully rational students could occur only in the presence of university's private information. To the extent that a university with active affirmative action programs is concerned about potential mismatch, it suggests that releasing more information to their applicants about how the admission officers feel about their fit with the university will minimize possibilities for actual mismatch. More transparency and more effective communication with the students, and possibly pre-enrollment sit-ins in college classrooms etc. can help minority students enrolling in an elite university only to find out that they would have been better off elsewhere.

We would also like to propose two potential avenues that may lead to a more conclusive test of mismatch effect.

The first potential avenue requires the cooperation of the selective university's Admission's Office. After the admission decisions are made, the Admission Officer (AO) could randomly group all the admitted minority students into two groups: the first group will receive the standard admission letter; and the second group will receive the standard admission letter together with additional information (e.g. the Admission Officer's evaluation rankings of the applicant) that the AO thinks could be relevant to predict the applicants' post-enrollment performance. Then if we observe that the enrollment rate for the second group is smaller than the first group, this will prove that the university's private information may have generated mismatch.

The second potential avenue to be more conclusive is to ask the admitted students two questions:

> **Q1.** "What do you realistically expect will be your cumulative GPA at Duke after your first year?"
>
> **Q2.** "To the extent that you may want to reconsider enrolling at this university if your cumulative GPA ends up being too low, how low should it be for you to change your mind about enrolling in this university?"

If a research with access to the Admission Officer's private information would have predicted a student's cumulative GPA to be lower than the stated threshold by the student in Q2, we could also conclude that there is mismatch.

7 Conclusion

We argue that once we take into account the students' rational enrollment decisions, mismatch in the sense that the intended beneficiary of affirmative action admission policies are made worse off could occur only if selective universities possess private information about students' post-enrollment treatment effects. This necessary condition for mismatch provides the basis for a new test. We propose an empirical methodology to test for private information in such a setting. The test is implemented using data from Campus Life and Learning Project (CLL) at Duke. Preliminary evidence shows that Duke does possess private information that is a statistically significant predictor of the students' post-enrollment academic performance, but Duke's private information only explains a very small percentage of the variation in student performance. We also propose strategies to evaluate more conclusively whether the evidence of Duke private information has generated mismatch.

References

[1] Arcidiacono, Peter and Jacob Vigdor (2007). "Does the River Spill Over? Estimating the Economic Returns to Attending a Racially Diverse College." Mimeo, Duke University.

[2] Arrow, Kenneth (1963). "Uncertainty and the Welfare Economics of Medicare Care." *American Economic Review,* 53 (December): 941-973.

[3] Ayres, Ian, and Richard Brooks (2005). "Does Affirmative Action Reduce the Number of Black Lawyers?" *Stanford Law Review*, Vol. 57, No. 6, 1807-1854.

[4] Barnes, Katherine Y. (2007). "Is Affirmative Action Responsible for the Achievement Gap Between Black and White Law Students?" *Northwestern University Law Review*, Vol. 101, No. 4, Fall: 1759-1808.

[5] Black, D., K. Daniel and J. Smith (2001). "Racial Differences in the Effects of College Quality and Student Body Diversity on Wages." in *Diversity Challenged*, Harvard Educational Review.

[6] Bryang, Anita-Yvonne, Kenneth I. Spenner and Nathan Martin, with Alexandra Rollins and Rebecca Tippett (2006). "The Campus Life and Learning Project: A Report on the First Two College Years." Available at http://www.soc.duke.edu/undergraduate/cll/final_report.pdf

[7] Bryang, Anita-Yvonne, Kenneth I. Spenner and Nathan Martin, with Jessica M. Sautter (2007). "The Campus Life and Learning Project: A Report on the College Career." Available at http://www.soc.duke.edu/undergraduate/cll/2nd_report.pdf

[8] Cardon, James H. and Igal Hendel (2001). "Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey." *Rand Journal of Economics,* Vol. 32 (Autumn): 408-427.

[9] Cawley, John and Thomas Philipson (1999). "An Empirical Examination of Information Barriers to Trade in Insurance." *American Economic Review,* Vol. 89, No. 5: 827-846.

[10] Chambers, David L., Timothy T. Clydesdale, William C. Kidder, and Richard O. Lempert (2005). "The Real Impact of Eliminating Affirmative Action in American Law Schools: An Empirical Critique of Richard Sander's Study." *Stanford Law Review*, Vol. 57, No. 6: 1855-1898.

[11] Chiappori, Pierre-André and Bernard Salanié (2000). "Testing for Asymmetric Information in Insurance Markets." *Journal of Political Economy,* Vol. 108, No. 2: 56-78.

[12] Chiappori, Pierre-André (2001). "Econometric Models of Insurance under Asymmetric Information." In *Handbook of Insurance*, edited by Georges Dionne. Springer.

[13] Chiappori, Pierre-André, Bruno Jullien, Bernard Salanié and Francois Salanié (2006). "Asymmetric Information in Insurance: General Testable Implications." *Rand Journal of Economics*, Vol. 37 (Winter): 783-798.

[14] Cunha, Flavio, James J. Heckman and Salvador Navarro (2005). "Separating Uncertainty from Heterogeneity in Life Cycle Earnings." *Oxford Economic Papers* (2004 Hicks Lecture), Vol. 57, No. 2, 191-261.

[15] Cureton, Edward E. and Ralph B. D'Agostino (1983). *Factor Analysis: An Applied Approach.* Lawrence Erlbaum Associates, Inc. Publishers: New Jersey.

[16] Dale, Stacy Berg, and Alan B. Krueger (2002). "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables." *Quarterly Journal of Economics,* Vol. 117, No. 4: 1491-1527.

[17] Duncan, G.J., J. Boisjoly, D.M. Levy, M. Kremer, and J. Eccles (2006). "Empathy or Antipathy? The Impact of Diversity." *American Economic Review*, Vol. 96, No. 6, pp.1890-1905.

[18] Fang, Hanming, Michael P. Keane and Dan Silverman (2008). "Sources of Advantageous Selection: Evidence from the Medigap Insurance Market." *Journal of Political Economy,* Vol. 116, No. 2, 303-350.

[19] Finkelstein, Amy and Kathleen McGarry (2006). "Multiple Dimensions of Private Information: Evidence from the Long-Term Care Insurance Market." *American Economic Review*, Vol. 96, No. 5: 938-958.

[20] Finkelstein, Amy and James Poterba (2004). "Adverse Selection in Insurance Markets: Policyholder Evidence from the U.K. Annuity Market." *Journal of Political Economy*, Vol. 112, No. 1: 183-208.

[21] Ho, Daniel E. (2005). "Why Affirmative Action Does Not Cause Black Students to Fail the Bar." *Yale Law Journal,* Vol. 114, No. 8, 1997-2004.

[22] Horowitz, Joel (1998). *Semiparametric Methods in Econometrics.* Springer.

[23] Kellough, J. Edward (2006). *Understanding Affirmative Action: Politics, Discrimination and the Search for Justice.* Georgetown University Press: Washtington D.C.

[24] Kim, Jae-On and Charles W. Mueller (1994). "Introduction to Factor Analysis: What It Is and How to Do It." in Michael S. Lewis-Beck ed., *Factor Analysis and Related Techniques* (International Handbooks of Quantitative Applications in the Social Sciences, Vol. 5): Sage Publications.

[25] Kotlarski, Ignacy (1967). "On Characterizing the Gamma and Normal Distribution." *Pacific Journal of Mathematics*, 20, 729-738.

[26] Krasnokutskaya, Elena (2004). "Identification and Estimation in Highway Procurement Auctions Under Unobserved Auction heterogeneity." mimeo, University of Pennsylvania.

[27] Li, Tong and Q. Vuong (1998). "Nonparametric Estimation of Measurement Error Model Using Multiple Indicators." *Journal of Multivariate Analysis,* Vol 65, 135-169.

[28] Pauly, Mark V. (1974). "Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection." *Quartely Journal of Economics,* Vol. 88, No. 1: 44-62.

[29] Rao, B.L.S. Prakasa (1992). *Identifiability in Stochastic Models: Characterization of Probability Distributions.* Academic Press: New York.

[30] Rothschild, Michael and Joseph E. Stiglitz (1976). "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Quarterly Journal of Economics,* 90 (November): 629-649.

[31] Rothstein, Jesse and Albert Yoon (2008). "Mismatch in Law School." mimeo, Princeton University.

[32] Sander, Richard H. (2004). "A Systemic Analysis of Affirmative Action in American Law Schools." *Stanford Law Review*, Vol. 57, No. 2, 367-483.

[33] Sander, Richard H. (2005a). "Mismeasuring the Mismatch: A Response to Ho." *Yale Law Journal*, Vol. 114, No. 8: 2005-2010.

[34] Sander, Richard H. (2005b). "Reply: A Reply to Critics." *Stanford Law Review*, Vol. 57, No. 6, 1963-2016.

[35] Wilson, Charles (1977). "A Model of Insurance Markets with Incomplete Information." *Journal of Economic Theory* 16 (December): 167-207.