NBER WORKING PAPER SERIES

DOES AFFIRMATIVE ACTION LEAD TO MISMATCH? A NEW TEST AND EVIDENCE

Peter Arcidiacono
Esteban M. Aucejo
Hanming Fang
Kenneth I. Spenner

Working Paper 14885
http://www.nber.org/papers/w14885

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2009

We would like to thank Judy Chevalier, Joe Hotz, Caroline Hoxby, Jon Levin, Jim Levinsohn, Tong Li, Jeff Smith, Justin Wolfers and seminar participants at Brown, South Carolina, Vanderbilt, Penn and NBER Public Economics and Higher Education Meetings for helpful comments. The authors gratefully acknowledge support for this research provided by grants from the Andrew W. Mellon Foundation and Duke University. All remaining errors are ours. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

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

Peter Arcidiacono, Esteban M. Aucejo, Hanming Fang, and Kenneth I. Spenner
NBER Working Paper No. 14885
April 2009
JEL No. D8,I28,J15

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. Evidence shows that Duke does possess private information that is a statistically significant predictor of the students' post-enrollment academic performance. We also propose strategies to evaluate more conclusively whether the evidence of Duke private information has generated mismatch.

Peter Arcidiacono
Department of Economics
201A Social Science
Duke University
Durham, NC 27708
and NBER
psarcidi@econ.duke.edu

Esteban M. Aucejo
Department of Economics
Duke University
213 Social Sciences Bldg
Box 90097
Durham, NC 27708-0097
esteban.aucejo@duke.edu

Hanming Fang
Department of Economics
Duke University
213 Social Sciences Building
Box 90097
Durham, NC 27708-0097
and NBER
hanming.fang@duke.edu

Kenneth I. Spenner
Department of Sociology
Duke University
Durham, NC 27708
kspen@soc.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 (2009), using information 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.[3] Duncan et. al. (2006), exploiting conditional 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 outcomes of minority students might have been worsened as a result of attending a selective university relative to attending a less selective school.

But even if some of the outcomes for minority students are worse under affirmative action, it still may be the case that minority students are better off under affirmative action. 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 just 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

[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 introduction to various arguments for and against affirmative action.

[3] Arcidiacono, Khan, and Vigdor (2008) also suggest that affirmative action actually leads to *less* inter-racial interaction due to the exacerbation of the within-school gap between minority and majority academic backgrounds.

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. 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.[4] 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 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 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. However, 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 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

[4] There is some evidence in the literature that students' expectations about their performance are inaccurate and updated over time. Stinebrickner and Stinebrickner (2008) have information at multiple points during the student's college career from Berea college. They find strong evidence of students updating their expectations over time and making decisions (such as the decision to drop out) based upon the new information they receive through their grades.

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.

We use 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 administered to all under-represented minorities in each of the cohorts as well as a random sample of whites and Asians. The CLL provides information about the participants' college expectations, social and family background, and satisfaction measures as well as providing confidential access to students academic 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.

We test whether Duke's private information is important to outcomes such as grade point average 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.

We will also discuss in Section 7 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.

The remainder of the paper is structured as follows. In Section 2 we discuss the mismatch literature. In Section 3 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 4 we describe the Campus Life and Learning (CLL) Project data that we use in our application

[5] Kotlarski theorem has been applied in economics in Krasnokutskaya (2008) and Cunha, Heckman and Navarro (2005).

to test for private information. In Section 5 we provide some baseline regressions to provide some preliminary bounds the importance of Duke and student private information in predicting students' performance at Duke. In Section 6 we describe a non-parametric empirical method to identify private information and present our main empirical results. In Section 7 we discuss two potential avenues to provide more conclusive evidence for mismatch; and Section 8 concludes. In the Appendix, we discuss some data attrition issues and report some omitted regression coefficients.

2 Mismatch Literature

The mismatch literature to date has focused on 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 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.[6] 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.[7] 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

[6] Loury and Garman (1995) appears to be the predecessor of the "mismatch" literature. They found that college selectivity and performance at college both have significant effects on earnings. The earnings gain by black students from attending selective colleges are offset by worse college performance for those Black students whose own SAT scores are significantly below the median of the college they attended, i.e. those "mismatched" blacks.

[7] 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.

selective schools are likely to have better unobserved credentials.[8] 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 provide 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, a group that makes up 25% of the sample. However, they could not conclusively find effects for the bottom 75% of the distribution due to not having enough whites with similar credentials. We will argue that the success of the top 25% is *necessary* for mismatch to occur if blacks at least know the overall relationship between credentials and success while only have expectations on their own credentials. Namely, if all blacks were mismatched then there would be no scope for students making rational decisions to attend schools where they were mismatched: there has to be some non-mismatched black students in order for rational mismatch to occur.

To summarize, the existing literature on the mismatch effect differs in the empirical strategy used to assess the counterfactual outcome of minority students attending less selective universities; and the evidence is mixed. We want to recast the mismatch problem in the context of rational decision making which, as show in the next section, points us towards examining whether universities have private information on the future success of their students.

3 The Model

Consider two universities that differ in selectiveness. For convenience, suppose that only one university is selective, which we refer to as the elite university. The elite university has an enrollment capacity 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.

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

[8] 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. A potential problem 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 clear that such a strategy does control for unobserved credentials.

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 university's characteristics. 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 a capacity constraint and to a diversity constraint.[9, 10] We assume that the student is risk neutral, and thus will choose the university (if she is admitted) that offers her the highest treatment effect.

3.1 The Case of Symmetric Information and Diversity Concerns

We first consider the case that the students know their treatment effects from attending the elite university.[11] In this symmetric information case, no students with a negative treatment effect will matriculate in the elite university, even if they are admitted. Note if the university admits a student of race r and treatment effect T_r', then it will also admit all students of race r who have treatment effects above T_r'. Let T_r^* denote the lowest treatment effect among those who are admitted of race r. Thus the matriculation constraint for the students must be

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

It is this constraint that effectively makes *ex ante* mismatch under symmetric information impossible: no student will attend a school where their treatment effect is negative. Note that the matriculation constraint must hold regardless of the objective function of the elite university.

The elite university's problem is then to maximize the treatment effect of its student body subject to three constraints:

- That enrollment is no larger than C (capacity constraint);
- That the fraction of blacks attending is no less than $\lambda \geq 0$ (diversity constraint);

[9] The elite university may have other factor besides the treatment of the student in their objection function such as future donations or whether the treatment effect is positive relative to not attending college. We note the effect of different objective functions throughout this section, though fundamentally a different objective function by the university will not change our conclusion that mismatch can *only* occur when the university has private information about the treatment effect for the student. It will become clear that the key driver of our result is *the rational matriculation constraint of the students, not* the objective function of the elite university.

[10] While we treat diversity as a constraint that must be satisfied here, the qualitative results do not change if we put a penalty function that penalizes deviations from optimal diversity levels into the objective function. These results are available upon request.

[11] Alternatively, both the student and the school could be uncertain about the treatment effect. The key assumption is that they are operating with the same information set: the university does not have private information.

- That the expected treatment effect for individuals of both races is positive (matriculation constraint).

The maximization problem is then:

$$\max_{\{T_w^*, T_b^*\}} \sum_{r \in \{w,b\}} N_r \int_{T_r^*}^{\infty} T_r f_r(T_r)\, dT_r \tag{1}$$

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

$$\frac{N_b [1 - F_b(T_b^*)]}{N_w [1 - F_w(T_w^*)]} \geq \frac{\lambda}{1 - \lambda}, \tag{3}$$

$$T_r^* \geq 0 \text{ for } r \in \{w, b\}, \tag{4}$$

We index the solutions to the above problem by $T_r^*(\lambda)$. Thus, the solution when there is no diversity constraint is $T_r^* = T_r^*(0)$. When $\lambda = 0$, the university is indifferent between black and white students conditional on their treatment effect, implying that $T_r^*(0) = T^*(0)$ for all r. If setting the cutoff treatment effect to zero does not violate the capacity constraint, then $T^*(0) = 0$. Otherwise, $T^*(0)$ uniquely solves

$$N_w [1 - F_w(T^*)] + N_b [1 - F_b(T^*)] = C.$$

Note that even though the admission cutoffs are the same for blacks and whites, the racial composition of the student body may be very different from the overall composition of the applicants because $F_w(\cdot) \neq F_b(\cdot)$.

The solution found when $\lambda = 0$ will be the same as the solution for all λ's that are sufficiently small. Denote as $p^*(0)$ the fraction of blacks in the student body when $\lambda = 0$:

$$p^*(0) = \frac{N_b [1 - F_b(T_b^*(0))]}{C}.$$

Let $\lambda_1 = p^*(0)$. If $\lambda < \lambda_1$, then the presence of the diversity constraint does not affect the solution to the elite university's maximization problem: varying λ in this range has no impact on the diversity of the elite university.

The second relevant cutoff point is when the admissions standard is set to zero, leading all blacks who have positive treatment effects to be admitted. Denote this cutoff by λ_2:

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

When $\lambda \in [\lambda_1, \lambda_2]$, the solution to the elite university's problem are implicitly characterized by:

$$N_b [1 - F_b(T_b^*(\lambda))] = \lambda C$$

$$N_w [1 - F_w(T_w^*(\lambda))] = (1 - \lambda) C.$$

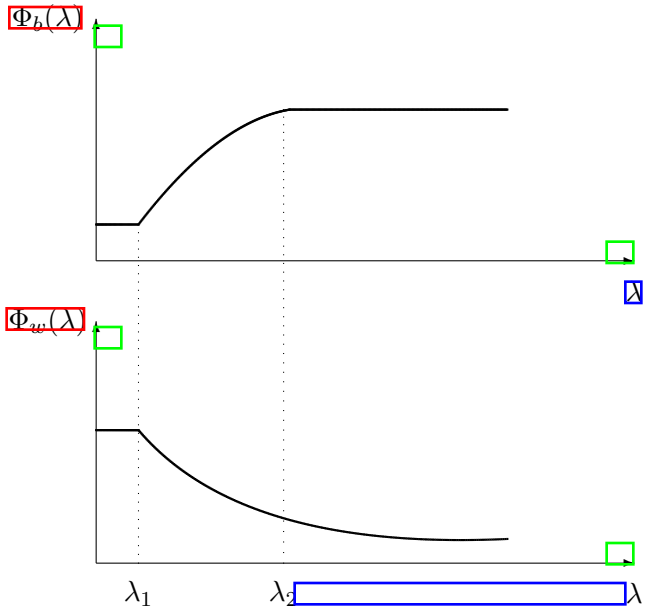

Figure 1: The Total Treatment Effects as a Function of the Diversity Concern λ: The Symmetric Information Case.

Notes: At $\lambda < \lambda_1$, the diversity constraint does not bind. At $\lambda > \lambda_2$, all blacks with positive treatment effects attend the elite university

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. No more blacks are induced to attend by increasing λ as all blacks who have positive treatment effects are already attending. The cutoff treatment effect for white, $T_w^*(\lambda)$, is then chosen so that

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

That is

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

Under this admission policy, the total enrollment is given by

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

which is less than the allowable capacity C.

We now have all the pieces we need to qualitative describe the total treatment effect for each race as a function of λ. Define the total treatment effect for group r as:

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

Given the above discussion, we know that $\Phi_r(\lambda)$ can be depicted as in Figure 1. Between 0 and λ_1, the treatment effect for blacks and whites is unchanged with increases in λ as the diversity constraint is slack. Between λ_1 and λ_2, the treatment effect for blacks rises at the expense of the treatment effect for whites. Past λ_2, the treatment effect for whites falls with no change in the black treatment effect as blacks are already at their maximum treatment effect at λ_2.

In sum, when both the university and the student operate from the same information set, diversity constraints at least weakly increase the treatment effect for blacks:

Proposition 1 *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.*

3.1.1 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. The elite university's optimization problem becomes:

$$\max_{\{T_w^*, T_b^*\}} \sum_{r \in \{w,b\}} N_r \int_{T_r^*}^{\infty} T_r f_r(T_r)\, dT_r \tag{5}$$

$$s.t. \quad \sum_{r \in \{w,b\}} N_r \left[1 - F_r(T_r^*)\right] \leq C, \tag{6}$$

$$\frac{N_b \left[1 - F_b(T_b^*)\right]}{N_w \left[1 - F_w(T_w^*)\right]} \geq \frac{\lambda}{1-\lambda}, \tag{7}$$

$$\mathrm{E}\left[T_r | T_r \geq T_r^*\right] \geq 0 \text{ for } r \in \{w, b\}, \tag{8}$$

where $\lambda \geq 0$ again measures the degree of the elite university's diversity concern. Note that the only difference between the case with asymmetric information from the case with symmetric information lies in the difference between the student matriculation constraints (4) and (8). 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 the elite university's maximization problem, 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

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

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 matriculation decisions. Note also that

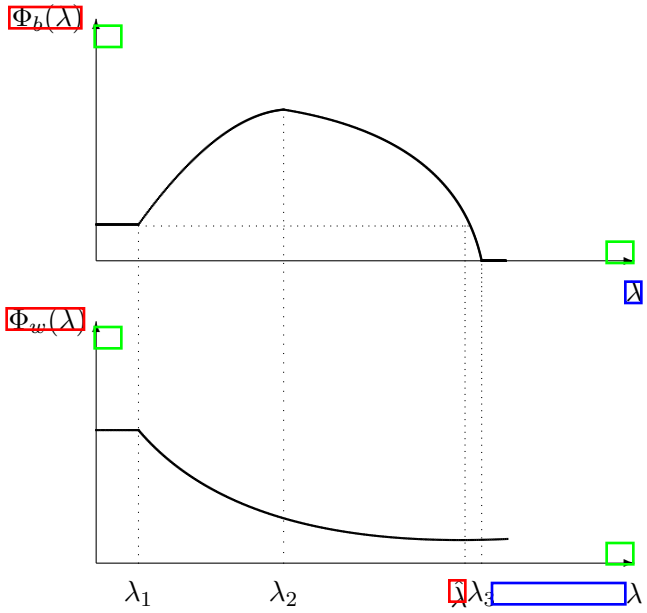

Figure 2: The Total Treatment Effects as a Function of the Diversity Concern λ: The Asymmetric Information Case.

Notes: At $\lambda \leq \lambda_1$, the diversity constraint does not bind. At $\lambda \geq \lambda_2$, the marginal admitted black has a negative treatment effect. At $\lambda > \lambda_3$, the matriculation constraint binds. For $\lambda \geq \hat{\lambda}$, blacks are overall worse off with affirmative action than without affirmative action in terms of *ex ante* expected treatment effects.

by definition, the total treatment effect for blacks at λ_3 is exactly zero:

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

Again consider the interesting case where the elite university's capacity constraint binds. The solution to the elite university's problem 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

$$N_b\left[1 - F_b\left(T_b^*(\lambda)\right)\right] = \lambda C$$

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

When $\lambda > \lambda_3$, the elite university can no longer increase black enrollment by lowering the admission standard because of the binding enrollment constraint (8). 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 2 *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 $\lambda > \hat{\lambda}$, $\Phi_b(\lambda) < \Phi_b(0)$.*

3.2 Mismatch and Asymmetric Information

We are now 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 blacks if some black students with negative treatment effects are admitted and enroll, 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 blacks if black students as a whole are made worse off in expectation, i.e.,*

$$\int_{T_b^*(\lambda)} T_b dF_b(T_b) < \int_{T_b^*(0)} T_b dF_b(T_b). \tag{9}$$

Equivalently, (9) can be written as

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

Note that $T_b^*(0) = T_b^* \geq 0$ 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 1 and 2, we have:

Proposition 3 *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.*

4 The Campus Life and Learning (CLL) Project Data

In Section 3, we argued that once we take into account the students' rational matriculation decisions, a *necessary* condition for either local mismatch or global mismatch to arise is that the elite university has private information about the students' treatment effects. In our empirical section, we propose tests for private information by the elite university. If our tests reject 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.

In this section, we describe data from the Campus Life and Learning Project (CLL) at Duke University that will allow us to test whether Duke has private information regarding the future success of their students.[12] 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).[13] 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.

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; with 91 percent of these respondents providing signed release of their institutional records for the study. In the spring semester of the

[12] 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).

[13] 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.

first, second and fourth college year, each cohort was again surveyed by mail.[14] However, response rates declined in the years following enrollment with 71, 65 and 59 percent responding in the first, second and fourth years of college, respectively.[15]

The pre-college survey provides detailed measurement of the students' social and family background, prior school experiences, social networks, and expectations of their college performance. In particular, students were asked

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

We can then relate this measure to the student's actual first year grade point average (GPA).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 who released access to their institutional records, we also have information about their grades, graduation outcomes, test scores (SAT and ACT) and financial aid and support. Further, we have the Duke Admission Officers' rankings of their applications on six measures: achievement, curriculum, essay, personal qualities, recommendations and test scores. Each of these rankings are reported on a five point scale. It is these rankings coupled with student expected performance that will be used to disentangle what the student knows from what the institution knows about how well the student will perform in college.

Table 1 contains summary statistics for the key variables in the CLL data set by race. The first rows reveal that there is substantial amount of variation in entering credentials among students of difference races. Asians and Whites tend to have higher evaluations by Duke Admission Officers in all six categories than black and Latino students, with test score showing by far the largest gap. Despite these differences in credentials, black and white students have quite similar expectations about their GPA during their first year in college (3.51 for whites and 3.44 for blacks).[16]

However, Table 1 shows that there is a significant racial difference in the actual first year cumulative GPAs. 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 Asians and whites have

[14] The survey was not conducted in the third year as many Duke students study abroad during that year.

[15] In the appendix we examine who attrits and test for non-repsonse bias.

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

substantially higher (more than one standard deviation) 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.

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 help these students form 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.

5 Baseline Regressions

While the CLL data set has the advantage of reporting both information from the students regarding their expected grades and information from Duke regarding the ranking of the applicant, disentangling Duke's private information from what the student knows is challenging. We begin by running some baseline regressions which may *bound* the amount of private information both the student and Duke have about student's performance.

We begin by examining the difference between the student's expected GPA for their freshman year, ExpGPA, and their actual cumulative GPA for their freshman year, GPA.[17] Specifically we see how forecastable this difference is with variables the student should know the effects of, such as their race and SAT scores. Let $\mathbf{Z}$ indicate this set of variables. We then add variables the student might only have partial information about such as Duke's ranking of the student (DukeEv). The forecast error for student i is then:

$$\text{GPA}_i - \text{ExpGPA}_i = \mathbf{Z}_i\boldsymbol{\alpha}_1 + \epsilon_{i1} \quad (10)$$

$$\text{GPA}_i - \text{ExpGPA}_i = \mathbf{Z}_i\boldsymbol{\alpha}_2 + \text{DukeEv}_i\beta_2 + \epsilon_{i2} \quad (11)$$

where the ϵ's are the projection errors.

Results from regressions (10) and (11) are reported in Table 2.[18] For ease of interpretation, we adjusted the SAT score such that it has zero mean and a standard deviation of one. Column 1 of Table 2 shows that students underestimate the relationship between their SAT score and performance.

[17] In our data, the correlation between student's actual cumulative GPA (GPA)and their expected GPA (ExpGPA)at the end of their first year is 0.178.

[18] We have also experimented with specifications that include high school characteristics (private, public, religious etc.) in the regressions. Their coefficients are not significant and they neither affect the other coefficient estimates, nor significantly increase the R^2 of the regressions.

Table 2: The Components of Students' Forecasting Error

Variable	(1)	(2)
Constant	-0.256**	-0.883***
	(0.046)	(0.370)
Male	-0.120***	-0.093***
	(0.037)	(0.036)
Black	-0.131**	-0.110*
	(0.060)	(0.063)
White	0.144***	0.118**
	(0.048)	(0.051)
Asian	-0.010	-0.051
	(0.057)	(0.059)
Adjusted SAT	0.106***	0.061***
	(0.022)	(0.023)
Controls for Duke Eval?	No	Yes
R^2	0.088	0.148

Notes: Dependent variables is (GPA−ExpGPA); N = 938. Adjusted SAT is the SAT score normalized to have zero mean and a standard deviation of one. The coefficients on the Duke evaluation rankings are reported in Table A.1 in the Appendix. *, ** and *** indicate that the coefficient is significant at 10%, 5% and 1% respectively.

Virtually all groups on average over-predict their performance, with the one exception being white females with SAT scores more than one standard deviation above the mean. As expected given the descriptive statistics in Table 1, blacks significantly overestimate their performance relative to the other racial groups. Further, the variance of (GPA-ExpGPA) is 0.27 and is actually *higher* than the variance of first year GPA, which is 0.22. Clearly if we assume that the student's only information about their future performance is captured in their expected GPA, then there is a lot of information that the university possesses and a significant amount of noise in expected GPA. Moreover, the statistically significant coefficient estimates on the Adjusted SAT and race variables indicates that the student does not accurately know how these characteristics translate into their future performance. Duke, however, is likely to know more accurately about the relationship between characteristics and performance. Column 2 in Table 2 adds controls for Duke's evaluation rankings of the students. The R^2 increases from 0.088 to 0.148 when we include Duke's rankings, again suggesting that Duke has either private information about the student's future performance or in how information known to both the student and Duke translates into future performance.[19]

[19] The coefficients on the Duke evaluation ranking variables are given in Table A.1 in the Appendix.

The expected GPA of the student, however, may not reflect the student's true information set. We now test whether the university has private information under a more restrictive setting. Namely, we assume students know how their SAT scores and other demographics translate into future performance. The information set for both the student and the university then contains this common observed information plus common information that is unobserved to the researcher. Under these assumptions, running the regression

$$\text{GPA}_i = \mathbf{Z}_i\boldsymbol{\alpha}_3 + \epsilon_{i3}, \tag{12}$$

and calculating the R^2 then leads to a lower bound on the amount of common information that the student and the university have regarding the student's future performance as it does not include common unobserved information. Results from this regression are reported Column 1 of Table 3.[20] Close to 19 percent of the variation in grades can be explained by these observables. Comparing this result with that in Column 1 of Table 1 suggests that students underestimate the relationship between SAT scores and performance by more than 50 percent.

To this baseline regression, we add the student's expected GPA:

$$\text{GPA}_i = \mathbf{Z}_i\boldsymbol{\alpha}_4 + \textsc{ExpGPA}_i\delta_4 + \epsilon_{i4}. \tag{13}$$

The difference in R^2 between (12) and (13) should provide an *upper* bound on the student's private information as it includes not only the student's private information, but also common unobserved information that is correlated with student's private information. These results are reported in Column 2 of Table 3. The differences in R^2 between Column 2 and Column 1 in Table 3 indicates that including the expected GPA of the student increases the R^2 by less than 0.01, which provides an upper bound of the importance of student's private information.

Finally, we add Duke's evaluation rankings of the students:

$$\text{GPA}_i = \mathbf{Z}_i\boldsymbol{\alpha}_5 + \textsc{ExpGPA}_i\delta_5 + \textsc{DukeEv}_i\boldsymbol{\beta}_5 + \epsilon_{i5}. \tag{14}$$

The difference in R^2 between (13) and (14) should provide a *lower* bound on the importance of Duke's private information. Notice from Column 3 that controlling for Duke's rankings increase the R^2 by more than 0.12, again suggesting substantial Duke private information.[21] Note that this still leaves two-thirds of the variation in GPA unexplained, perhaps due to course selection and shocks to how students respond to college life.

[20] Adding additional variables such as family income and mother's education had little effect on the R^2 but did lead to some attrition.

[21] One can also reverse the order of the regressions such that we first control for Duke's evaluation rankings and then add student's expected GPA. The addition of the student's expected GPA in this order increases the R^2 by only 0.001, with an insignificant coefficient estimate on expected GPA.

Table 3: Baseline Tests of Private Information

Variable	(1)	(2)	(3)
Constant	3.309***	2.792***	1.828***
	(0.039)	(0.187)	(0.325)
Male	-0.080**	-0.086***	-0.043
	(0.037)	(0.036)	(0.029)
Black	-0.191***	-0.182***	-0.158***
	(0.051)	(0.052)	(0.053)
White	0.047	0.061	0.030
	(0.041)	(0.041)	(0.042)
Asian	0.037	0.030	-0.010
	(0.049)	(0.049)	(0.049)
Adjusted SAT	0.178***	0.167***	0.103***
	(0.018)	(0.018)	(0.017)
Expected GPA		0.145***	0.050
		(0.052)	(0.047)
Controls for Duke Eval?	No	No	Yes
R^2	0.188	0.196	0.321

Notes: Dependent variables is GPA; N = 938. Adjusted SAT is the SAT score normalized to have zero mean and a standard deviation of one. The coefficients on the Duke evaluation rankings are reported in Table A.1 in the Appendix. *, ** and *** indicate that the coefficient is significant at 10%, 5% and 1% respectively.

A drawback of this empirical strategy is that we do not fully observe common information; as a consequence, there is no guarantee that the reported bounds are in fact the real ones. Further, measurement error in the expected GPA variable may be contaminating the results. In the following section, we implement a different strategy that overcomes these limitations; and it allows us to identify private and common information in order to perform a more accurate variance decomposition analysis. However, it is worth mentioning that both strategies provide surprisingly similar results.

6 Non-Parametric Identification of Private Information

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.[22] 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).[23]

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 of the market, the potential insured, and it is manifested through their insurance purchase and their *ex post* risk occurrence. In our setting, there is presumably 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 describe below a new empirical strategy to identify private information in our setting.

6.1 Available Data and Assumptions

As we mention in section 3, we have data about an **observed student outcome** Y (i.e. first year cumulative GPA, denoted by GPA). Conceptually, we assume that Y is a linear function of

[22] 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.

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

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 $\mathbf{Z}$ 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 $\mathbf{Z}$ for the discussion here for simplicity.

Specifically, suppose that

$$Y = X_C\gamma_C + X_U\gamma_U + X_S\gamma_S + \varepsilon, \tag{15}$$

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.

Suppose 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.[24] We assume that (W_U, W_S) are related to X_C, X_U and X_S as follows:

$$W_U = X_C + X_U, \tag{16}$$

$$W_S = X_C + X_S. \tag{17}$$

To summarize, suppose that we observe a data set consisting $\{W_U, W_S, Y\}$ and assume that there exists independent variables X_C, X_U, X_S and ε such that $\{W_U, W_S, Y\}$ are generated by (15)-(17).

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 Y in the data?

6.2 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 $\{X_{Ci}, X_{Ui}, X_{Si}\}$ from the marginal distributions of X_C, X_U and X_S recovered in step 1;

[24] We will describe in Subsection 6.3 below the empirical counterparts of W_U and W_S in our setting.

3. We obtain samples of $\{W_{Ui}, W_{Si}\}$ from the random samples of $\{X_{Ci}, X_{Ui}, X_{Si}\}$ generated in step two, and then recover a sample of Y_i conditional on $\{W_{Ui}, W_{Si}\}$ using multiple imputation methods.[25]

4. We run regressions of Y on X_C, X_U, X_S using the pseudo-sample $\{Y_i, X_{Ci}, X_{Ui}, X_{Si}\}$ simulated above to estimate γ_C, γ_U and γ_S, and to do variance decomposition.

Now we provide more details about the above empirical strategy. The key is the first step which uses a mathematical result known as 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 (16) and (17). 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).[26] The proof of the theorem also suggests how the marginal distributions for X_C, X_S and X_U can be constructed. Let

$$\Psi(t_1, t_2) = \mathrm{E} \exp(it_1 W_U + it_2 W_S) \tag{18}$$

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

$$\begin{aligned} \Psi_1(t_1, t_2) &\equiv \frac{\partial \Psi(t_1, t_2)}{\partial t_1} \\ &= \mathrm{E}[iW_U \exp(it_1 W_U + it_2 W_S)] \end{aligned} \tag{19}$$

denote the derivative of $\Psi(\cdot, \cdot)$ 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

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

[25] See Rubin (1987) for an extensive description of this methodology.

[26] 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 (2008) 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.

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)

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

Once we have the marginal distributions for X_K for $K \in \{C,U,S\}$, the remaining steps 2-4 described above are rather straightforward. Now we describe the somewhat standard estimation procedure to carry out step 1.[27] The key is to estimate $\Psi(\cdot,\cdot)$ and $\Psi_1(\cdot,,\cdot)$ by their sample analogs: given a sample $\left\{W_U^j, W_S^j\right\}_{j=1}^n$,

$$\widehat{\Psi(t_1,t_2)} = \frac{1}{n}\sum_{j=1}^{n} \exp\left(it_1W_U^j + it_2W_S^j\right)$$

$$\widehat{\Psi_1(t_1,t_2)} = \frac{1}{n}\sum_{j=1}^{n} iW_U^j \exp\left(it_1W_U^j + it_2W_S^j\right),$$

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

Remarks. We have assumed in equation (15) that the student outcome Y 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 Y as a nonlinear function of these variables, or even non-parametrically estimate their relations.

It is also worth noting in specification (16) and (17), 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 of the 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 γ_U and γ_S downward, but when we do variance decomposition for Y, 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 Y.

6.3 Implementation Details and Results

As we have already mentioned, it is necessary to have access to (at least) two variables $\{W_U, W_S\}$ in order to apply Kotlarski's decomposition. Here we provide the details of these variables in our empirical application.

[27] See Krasnokutskaya (2008) for similar estimation procedure. Horowitz (1998, Chapter 4) describes some useful suggestions for issues related to smoothing.

W_U is specified as Duke's predicted first year GPA for the student, which we denote by GPA_U. Specifically, GPA_U is predicted student GPA from the estimated regression

$$\text{GPA}_i = \mathbf{Z}_i\boldsymbol{\alpha}_6 + \text{DUKEEV}_i\boldsymbol{\beta}_6 + \epsilon_{i6},$$

where GPA_i denotes the actual first year GPA. Recall that $\mathbf{Z}_i$ are the observed SAT scores and demographics and DUKEEV refers to the Duke ranking variables.

For W_S, we consider two alternative specifications. The first specification for W_S is the student's predicted GPA, which we denote by GPA_S, predicted from the estimated regression equation (13) from the previous section:

$$\text{GPA}_i = Z_i\alpha_4 + \text{EXPGPA}_i\beta_4 + \epsilon_{i4}.$$

This specification implies that students have an accurate idea about how to weight each informational variable (e.g. SAT) when they predict their performance. The second specification for W_S is the expected GPA (EXPGPA) reported by the student before coming to Duke in the CLL survey. To the extent that the students may not properly weigh the effect of the observable variables on their actual GPA, as documented in Table 2, we will be attributing some of the students' wrong weighting on the importance of common information X_C to Duke private information.

Applying Kotlarski's decomposition to $\{W_U, W_S\}$ allow us to recover a sample of $\{X_{Ci}, X_{Ui}, X_{Si}\}$, and to construct a sample of $\{W_{Ui}, W_{Si}\}$. The next step is to obtain a sample of grades (i.e. Y_i) conditional on W_{Ui} and W_{Si} by multiple imputation, which we follow Rubin (1987).[28, 29]

Once we have Y_i and $\{X_{Ci}, X_{Ui}, X_{Si}\}$, we perform a variance decomposition analysis (keeping in mind that X_{Ci}, X_{Ui}, X_{Si} are orthogonal to each other) to establish the contributions of Duke and students private and common information to the variation in GPA.

Table 4 reports the variance decomposition of GPA following two different specifications for W_S as described above. Specification (1) assumes that students know how to weight the available information when they predict their performance; results show that Duke's private information explains 9.1 percent of the variance in the students' actual first year cumulative GPA; the student's private information explains no more than 0.05 percent and the common information 26.5

[28] The basic steps of Rubin multiple imputation are as follows. (1). Calculate $V = (W'W)^{-1}$, $\widehat{\beta} = VW'Y$ and $\widehat{Y} = W'\widehat{\beta}$ where $W = \{W_U, W_S\}$; (2). Draw a random g from χ^2 distribution with degree of freedom $n_{obs} - r$; (3). Calculate $\sigma_*^2 = (Y - \widehat{Y})'(Y - \widehat{Y})/g$; (4). Draw an r-dimensional Normal random vector $D \tilde{} N(0, I_r)$, where I_r is the identity matrix of order r; (5). Calculate $\beta_* = \widehat{\beta} + \sigma V^{1/2} D$, where $V^{1/2}$ is the triangular square root of V obtained by the Cholesky decomposition; (6). Calculate predicted values $\widehat{Y}_i = W_i'\widehat{\beta}_*$; (7). For each missing value find the respondent whose $\widehat{Y}$ is closest to $\widehat{Y}_i$ and take Y of this respondent as the imputed value (predictive mean matching).

[29] In order to test for robustness of the results we also implemented a nonparametric approach to recover Y_i. Basically, we draw a sample of Z_i conditional on $\{W_{Ui}, W_{Si}\}$ from the observed conditional distribution $G(Y|W_U, W_S)$, which was obtained using the Epanechnikov kernel ($K(u) = \frac{3}{4}(1-u^2)1_{(|u|\leq 1)}$). The smoothing parameter was selected by following a refined plug in method, which tries to find the bandwidth that minimizes the mean integrated square error. Results obtained using this strategy did not differ significantly from those using multiple imputation technique.

Table 4: Regressing GPA on Duke Private Information, Student Private Information and Common Information.

	(1)			(2)		
	$W_U = \text{GPA}_U, W_S = \text{GPA}_S$			$W_U = \text{GPA}_U, W_S = \text{ExpGPA}$		
	Coef.	Std. Err.	R^2	Coef.	Std. Err.	R^2
Duke Priv. Inf. (X_U)	1.070***	0.037	0.091	0.957***	0.023	0.235
Student Priv. Inf. (X_S)	0.066*	0.040	0.0004	0.037*	0.022	0.0005
Common Inf. (X_C)	0.993***	0.018	0.265	0.994***	0.044	0.081
Total			0.356			0.317

Notes: *, ** and *** indicate that the coefficient is significant at 10%, 5% and 1% respectively.

percent. Specification (2) allows that students may not know how to weight the information, as a consequence, the fraction of the variance in GPA explained by Duke private information increases to 23.5 percent, that by common information declines to 8.1 percent, but the fraction explained by student private information remains about 0.1 percent.

It is worth noting the changes in R^2 depending on the specification. First the total R^2 in specification (2) is smaller than in specification (1), this could be due to the loss of valuable information when students do not correctly weight the available information or it could be due to students reporting expected GPA with error. Second, there is an important change, similar in magnitudes but in opposite directions, of the proportion of the variance that could be explain by common information and duke private information. This seems to suggest that the size of Duke private information not only depends on what information is not available to the students, but also how they weigh the information available to them in forecasting their performance at Duke.

Assuming that students are rational implies that coefficient on students private information should be equal one; however as we can see from Table 4, this is not the case. One possible explanation to this discrepancy is that students may report with error their expected GPA. The attenuation bias from the measurement error might drive the small R^2 we found for the students private information reported in Table 4. However, if we assume that the discrepancy between estimated $\hat{\gamma}_S$ and the postulated value $\gamma_S = 1$ under rational expectations is completely due to measurement error, we can easily provide an estimate of the variance of the student private information without measurement error. To see this, note that in the case of orthogonal explanatory variables with classical errors-in-variables, we have:

$$\hat{\gamma}_S = \gamma_S \frac{\text{Var}(X_S^*)}{\text{Var}(X_S)},$$

where $\text{Var}(X_S^*)$ is the variance of the student private information when it is purged of measurement error, and $\text{Var}(X_S)$ is X_S^* measured with error. Given that we know $\hat{\gamma}_S, \gamma_S$ (which is equal 1

under the rational expectation assumption) and $\text{Var}(X_S)$, then $\text{Var}(X_S^*) = \widehat{\gamma}_S \text{Var}(X_S)$. Thus the fraction of the variation in GPA explained by X_S^*, denoted by R^{2*}, is simply the R^2 reported in Table 4 divided by $\widehat{\gamma}_S$. Therefore, once we correct for measurement error, the fraction of the variation in GPA that is explained by student private information measured without error (i.e. X_S^*) under specifications (1) and (2) are respectively equal to 0.006 ($\approx$ 0.0004/0.066) and 0.0135 ($\approx$ 0.0005/0.037); again, both are substantially smaller in magnitude than the private information possessed by Duke.

Finally, the results obtained in this section are quite similar to those obtained from the baseline regressions. Thus, the conclusion that Duke does possess private information that can predict the students' post-enrollment performance is robust to different empirical strategies.

7 Discussion

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, even though we have shown substantial evidence that Duke does possess private information about the student's future performance, we can not conclude that there is mismatch.

We would also like to propose two potential avenues that may lead to a more conclusive test of mismatch. The first potential avenue requires the cooperation of the selective university's Admission's Office. After the admission decisions are made, the Admissions Officer could randomly assign 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 Admissions Officer's evaluation rankings of the applicant) that the Admissions Officer thinks are 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 test for mismatch 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.** "Suppose your expected GPA at Duke was X. Would you still have chosen Duke?"

If a researcher 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.[30]

[30] Note that in both cases we would be testing for local mismatch rather than global mismatch.

However, it is worth noting that even if one cannot conclusively prove the existence of mismatch, evidence that a selective university possesses valuable *ex ante* information could be used in *preventing* mismatch. 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 potentially find out that they would have been better off elsewhere.

8 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. The evidence shows that Duke does possess private information that is a statistically significant predictor of the students' post-enrollment academic performance. We also propose strategies to evaluate more conclusively whether the evidence of Duke private information has generated mismatch.

Appendix.

In this appendix, we examine the CLL data for drop-out bias and non-response bias. Also, we report the coefficients for the Duke ranking measures from Tables 2 and 3.

A Drop-out Bias and Non-Response Bias

The Registrar's Office data provided information on students who were not enrolled at the end semester in each survey year. Non-enrollment might occur for multiple reasons including academic or disciplinary probation, medical or personal leave of absence, dismissal or voluntary (including a small number of transfers) or involuntary withdrawal. Fewer than one percent of students ($n = 12$) were not enrolled at the end of the first year; about three percent by the end of the second year ($n = 48$) and just over five percent ($n = 81$) by the end of the senior year. We combined all of these reasons and tested for differences in selected admissions file information of those enrolled versus not enrolled at the end of each survey year. The test variables included racial ethnic group, SAT verbal and mathematics score, high school rank (where available), overall admission rating (a composite of five different measures), parental education, financial aid applicant, public-private non-religious-private religious high school and US citizenship. Of over 40 statistical tests, only two produced significant differences (with p-value less than 0.05): (1). At the end of the first year, dropouts had SAT-verbal scores of 734 versus 680 for non-dropouts; (2). by the end of the fourth year, those who had left college had an overall admissions rating of 46.0 (on a 0-60 scale) while those in college had an average rating of 49.7. No other differences were significant. We conclude that our data contain very little drop-out bias.

We conducted similar tests for respondents versus non-respondents for each wave for the same variable set plus college major (in 4 categories: engineering, natural science/mathematics, social science, humanities), whether or not the student was a legacy admission, and GPA in the semester previous to the survey semester. Seven variables show no significant differences or only a few small sporadic differences (one wave but not others), including racial ethnic category, high school rank, admissions rating, legacy, citizenship, financial aid applicant, and major group. However, several other variables show more systematic differences:

- Non-respondents at every wave have lower SAT scores (math: 9-15 points lower, roughly one-tenth to one-fifth of a standard deviation; verbal: 18-22 points lower, roughly one-third of a standard deviation).

- Non-respondents have slightly better educated parents at waves one and three, but not waves two and four.

- Non-respondents at every wave are less likely to be from a public high school and somewhat more likely to be from a private (non-religious) high school.

- Non-respondents have somewhat lower GPA in the previous semester compared with respondents (by about one-quarter of a letter grade).

These differences are somewhat inconsistent in that they include lower SAT and GPA for non-respondents, but higher parental education and private (more expensive) high schools. In general, the non-response bias is largest in the pre-college wave and smaller in the in-college waves even though the largest response rates are in the pre-college wave. In general, we judge the non-response bias as relatively minor on most variables and perhaps modest on SAT measures.

B Omitted Coefficients For Duke Evaluation Rankings in Tables 2 and 3

Here we report coefficients for the Duke ranking variables that were omitted from Tables 2 and 3. Column 1 shows the coefficients when the dependent variable is GPA-ExpGPA, the omitted coefficients from column 2, Table 2. Column 2 shows the coefficients when the dependent variable is GPA, the omitted coefficients from column 3, Table 3. The Admission Officer's ranking of the student's achievement and personal qualities are very significant in both regressions suggesting that they may be the key variable for Duke's private information. Recommendations, however, are only significant in the second column, suggesting that student's may have some idea of the informational content of their recommendation letters.

Table A.1: Coefficients on Duke Evaluation Rankings

	(1)	(2)
	(GPA-ExpGPA)	GPA
Achievement_3	0.217	0.227**
	(0.137)	(0.103)
Achievement_4	0.256*	0.305***
	(0.138)	(0.105)
Achievement_5	0.448***	0.520***
	(0.135)	(0.102)
Curriculum_3	0.307	0.301
	(0.275)	(0.224)
Curriculum_4	0.246	0.400*
	(0.258)	(0.212)
Curriculum_5	0.273	0.452**
	(0.259)	(0.213)
Essay_3	-0.086	-0.104
	(0.105)	(0.103)
Essay_4	-0.026	-0.038
	(0.107)	(0.104)
Essay_5	-0.124	-0.196
	(0.137)	(0.129)
Personal Qualities_3	0.053	0.116
	(0.198)	(0.168)
Personal Qualities_4	0.047	0.118
	(0.198)	(0.168)
Personal Qualities_5	0.213	0.305
	(0.209	(0.175)
Recommendations_3	0.010	0.393**
	(0.210)	(0.168)
Recommendations_4	0.026	0.423**
	(0.217)	(0.173)
Recommendations_5	0.014	0.427**
	(0.221)	(0.176)

Notes: Base category for each evaluation measure is 2, none of the sample had 1's for any of these measures. Column 1 refers to the omitted coefficients in Table 2 (Column 2), Column 2 refers to the omitted coefficients in Table 3 (Column 3). *, ** and *** indicate that the coefficient is significant at 10%, 5% and 1% respectively.

References

[1] Arcidiacono, Peter and Jacob Vigdor (2009). "Does the River Spill Over? Estimating the Economic Returns to Attending a Racially Diverse College." forthcoming, *Economic Inquiry.*

[2] Arcidiacono, Peter, Shakeeb Khan, and Jacob Vigdor (2008). "Representation versus Assimilation: How do Preference in College Admissions Affect Social Interactions? " mimeo, Duke University.

[3] Arrow, Kenneth (1963). "Uncertainty and the Welfare Economics of Medicare Care." *American Economic Review,* Vol. 53, No. 6, 941-973.

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

[5] 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.

[6] 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.

[7] Bryant, 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

[8] Bryant, 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

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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.

[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: 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: Washington D.C.

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

[25] Krasnokutskaya, Elena (2008). "Identification and Estimation in Highway Procurement Auctions Under Unobserved Auction heterogeneity." forthcoming, *Review of Economic Studies*.

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

[27] Loury, Linda D. and David Garman (1995). "College Selectivity and Earnings." *Journal of Labor Economics*, Vol. 13, No. 2: 289-308.

[28] Pauly, Mark V. (1974). "Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection." *Quarterly 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] Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

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

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

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

[36] Stinebrickner, Todd and Ralph Stinebrickner (2008). "Learning about academic ability and the college drop-out decision." mimeo, University of Western Ontario.

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