Stanford Center for International Development

Working Paper No. 426

Affirmative Action through Quotas: The Effect on Learning In India

by

Anjini Kochar

October 2010

Stanford University
John A. and Cynthia Fry Gunn Building, 366 Galvez Street
Stanford, CA 94305-6015

Affirmative Action through Quotas: The Effect on Learning In India

Anjini Kochar
Stanford University

October 2010

Countries characterized by significant learning inequalities across racial and ethnic groups frequently address them through a quota system in higher education. Using data from India, this paper estimates the effect of quotas on learning, quantifying their cost by comparing outcomes to those obtained under alternative admission rules. It does so by assessing how learning is affected by the mean and variance of classroom ability. I use control functions to eliminate selection bias, identifying them through a novel strategy which exploits the fact that students choose programs based on the partial information available at enrollment, information that determines enrollment but not learning conditional on enrollment.

Keywords: quotas, learning, control functions, partial information
JEL codes: I21, I28, O15

Acknowledgements: I would like to thank the Principal of the college and the Dean of Students who made this data available to me, and spent many hours explaining the academic system in the college, as well as the operation of the affirmative action program. Financial support from Forbes Marshall Pvt. Ltd., India, is gratefully acknowledged. The views expressed in this paper are those of the author alone.

1. Introduction

Countries in which minority groups have suffered from historic discrimination are commonly characterized by considerable schooling inequality, measured in terms of learning or cognitive skills, between these groups and the majority of the population. These inequalities exist even *within* schooling levels, for example, within the population of students who have completed 12 years of schooling. In India, data from the 2007 entrance examination tests for engineering colleges in the state of Maharashtra reveal that 41% of students who scored in the lowest decile are from backward or "scheduled" castes and tribes, even though they represent only 10% of the approximately 15,000 students who took the examination. Scheduled caste students comprised only 2% of students in the top decile. The average test score for upper caste students in this examination was 147 (out of 200), but only 125 for students of scheduled castes and tribes.

India, along with countries such as Malaysia, South Africa and Brazil which are also characterized by significant schooling inequality, has sought to remedy the situation through strong affirmative action programs in higher education, eschewing voluntary preferential systems in favor of a system of quotas which reserve a fixed percentage of seats in higher education institutions for groups that are the subject of this policy. The magnitude of the pre-existing inequality in learning implies, however, that such a system places minority students in classrooms with majority students of considerably higher levels of ability, significantly enhancing the variance in ability levels within a classroom. This may inhibit learning, not only of minority students, but also of majority students. The potential of the system to reduce learning has, in fact, been a concern of policy makers. India's Mandal Commission, which argued in favor of extending the country's system of quotas for students of lower castes to cover Other Backward Castes, also emphasized the need to ensure effective learning, stating that the country should not be "landed with ill-equipped and sub-standard engineers, doctors and other professionals."

Scant research exists on the effect of quotas on learning. Because quotas exist primarily in countries characterized by considerable inequality in cognitive skills, they induce a variance in the ability of students within a classroom which far exceeds that which exists in other countries, including those which use a preferential system of affirmative action as in the United States. Consequently, evidence accumulated from other countries is unlikely to apply to countries which use quotas.

This study examines the effect of India's quota system on learning. I use data on three cohorts of students from one of the top engineering colleges in the state of Maharashtra. Maharashtra reserves as many as 50%

of seats in engineering colleges for students of backward castes. Within each caste grouping (including upper castes), 30% of seats are reserved for women. These quotas apply for each field of study offered by a college, with students applying for fields within colleges, rather than to the college itself. Each field of study therefore essentially operates as a separate college. The college I study offers 9 different programs.

To evaluate the effect of the quota system on learning and compare outcomes under alternative admission rules, I follow Arcidiacono (2005) by characterizing programs by attributes of the student population, and predicting college placements under alternative assumptions regarding admissions rules.[1] The attributes I consider are the mean and variance of the (initial) ability of students in the program, the features of student composition which are most affected by the affirmative action policy. In this regard, this research is closely related to the empirical literature which assesses the effect of peer composition on individual achievement (Hanushek, Kain, Markman and Rivkin 2003; Hanushek and Rivkin 2006; Ding and Lehrer 2007, Hoxby 2000).

A major difficulty in empirical research on this topic is the selectivity bias which arises because of unobservable factors which affect a college's decision of which students to admit as well as the student's choice of college to attend from amongst those he or she is admitted to, factors which are likely correlated with individual learning. For example, estimates of the effect of the mean ability of the student body on individual learning will be biased if a student's choice of a college, and hence of the composition of students in the college, is correlated with unobservable individual characteristics which directly affect the student's performance.

Studies on the topic have attempted to control for selectivity bias in a variety of ways, including the use of rich data which allow researchers to follow students over the course of their schooling careers or enable comparisons across students of similar ability who nevertheless attend schools which differ in student composition (Dale and Krueger 2002). Recent studies have also sought identification by exploiting conditions in systems in which admission decisions are based on a deterministic rule. An example is one in which students are "tracked" into different sections based on their academic performance, and placement does not reflect student choice. Duflo, Dupas and Kremer (2008), for example, exploit an experimental design by which some students within a school were randomly assigned to a tracked program while others were not. However, the limited amount of variation in schooling ability within tracked classrooms and, indeed, across classrooms in any regular schooling system,

[1] As fully described later, this exercise is made easier by the fact that college admission rules are based solely on how students do in college entrance examinations.

generates estimates of the effect of student composition on learning which are unlikely to apply to environments with far greater variation.

Deterministic admission rules are also a feature of meritocracies, where admission to college is based purely on the student's academic standing, as most commonly assessed by their score in an entrance examination. In such instances, the returns to admission in selective colleges can be estimated by using a regression discontinuity framework, exploiting the fact that observationally similar students above and below the cut-off entrance score are assigned to colleges which differ in their selectivity (Hoekstra 2009). Alternatively, the fact that placement is based only on observables, here the student's test score, implies that selection bias can be controlled for by including test scores amongst the regressors (Ding and Lehrer 2007).

Even in a meritocracy, however, student choice is likely to play some role in educational placements. Students can reject the college they are admitted to for a less selective one because of factors such as location (and hence costs) but perhaps also because of information they possess on attributes of the college and their assessment of the goodness of "fit" between the college and their own abilities and personalities. If eligibility or placement rules do not exclusively determine enrollment in a program, then a strong regression discontinuity framework no longer applies; students below any given cut-off score include those who choose the less selective program. Nevertheless, eligibility rules can still be used as the basis of an instrumental variable strategy to identify the effect of enrollment outcomes on learning or earnings, as in Bertrand, Hanna and Mullainathan (2010).

A reduced form analysis of the returns to a selective college does not, however, provide the information necessary to evaluate the current system relative to others organized under different admission rules. To do so, it is necessary to characterize colleges by attributes such as the mean and variance of ability in the student population and to assess the returns to these attributes. However, if a student's choice of college is based partly on their unobserved skill set, then the regression error term will be correlated with enrollment outcomes. Consequently, a traditional instrumental variable approach to identifying the effect of student composition on learning using determinants of enrollment rules as instruments will not address selection bias; these determinants will be correlated with the regression error term.

I use control functions to eliminate the dependence of the regression error term on the student's choice of a program (Heckman and Robb 1985, 1986). Control functions are in turn identified through a novel strategy which exploits the fact that enrollment decisions reflect students' expectations of the selectivity of other programs relative to the one under consideration and that expectations can only reflect the information students have at the time of making their enrollment choices to infer program selectivity. The information is partial; the actual composition of the student body will only be known to any given student at the onset of the program. This partial information set therefore satisfies necessary exclusion restrictions: It determines enrollment choices, but not academic outcomes subsequent to enrollment.

To preview the results of this paper, I find that all students are adversely affected by the variance of student ability within a classroom. Students differ, however, in their response to mean (initial) ability, with students at the top and bottom of the ability distribution within a classroom being most adversely affected by the affirmative action policy.

As a consequence, the benchmark case of a pure meritocracy, without any quotas, would improve learning for *all* castes, including scheduled castes and tribes. This suggests a significant cost of the current system, measured by the reduction in learning it generates for all students. This benchmark case is purely hypothetical; few would advocate a system characterized by the significant caste segregation a pure meritocracy, in the Indian context, would imply. However, an alternative program, similar to one currently being implemented in India's top engineering colleges, which sets a minimum admission standard for scheduled caste and tribe students equal to the lowest entrance score recorded by a student from Other Backward Castes also yields improvements in learning for all students. The improvement is similar in magnitude to that achieved under a pure meritocracy for upper caste students, and for students from Other Backward Castes and Scheduled castes, and exceeds the meritocracy benchmark for scheduled tribes.

The results suggest that the negative effects of the current policy on learning, measured by the difference in cumulative grade point average (CGPA) achieved under the current system relative to the meritocracy, could be eliminated if caste-based gaps in cognitive skills at lower levels of schooling can be reduced. A more effective affirmative action policy should thus target learning gaps at elementary and secondary schooling levels.

The rest of this paper is structured as follows. Section 2 describes admission and academic procedures in engineering colleges in Maharashtra, focusing on conditions in the college under study. The theoretical and empirical framework for the analysis of this paper is then described in sections 3 and 4 respectively. Section 5 presents summary statistics and simple regression evidence on caste gaps in cognitive skills. The main empirical results of this paper are discussed in section 6. Predictions regarding learning under alternative admission rules are in section 7. Section 8 concludes.

2. Data and Research Setting: Maharashtra's Engineering Colleges

2.1 The Maharashtra Government's Procedures for Engineering Colleges

The Maharashtra Government has adopted a more stringent system of quotas than exists in other states of India, with 50% of seats in engineering colleges being reserved for students of lower castes, including scheduled castes and tribes, other nomadic tribes, and other backward castes.[2] As previously discussed, the quota system applies not at the level of the college, but to each of the disciplines or courses offered within a college. Thus, each program or course offered within a college operates as a separate unit for admissions purposes; students apply to a course within a college, rather than to the college itself.

The system of reservations is implemented through a Centralized Admission Process (CAP) which regulates admission into all government and government aided engineering colleges, as well as to unaided (private) colleges.[3] Admissions are based exclusively on the student's score in a common entrance examination, the Maharashtra Common Entrance Test (MAH-CET), subject to the student having received an average of at least 50% marks in Physics, Chemistry and Mathematics in 12th standard examinations and in the CET.[4] In each college, and for each program, the total number of available seats is determined by the All India Council of Technical Education (AICTE) prior to the start of each academic year. Because AICTE stipulates a teacher-pupil ratio of 1:60, the vast majority of approved programs have an intake of 60 students. The course-specific "sanctioned intake" forms the basis for admissions decisions into each program in that year. Given the fixed number of seats, existing quotas for each caste category of students determine the total number of students under each category who can be admitted.

The allocation of students to different courses proceeds strictly on the basis of their ranking in the entrance examination, normally conducted in early May of each year. Merit lists based on examination results, which provide students' overall rank as well as their category ranks, are publicly posted. The merit list provides a unique rank to each student, breaking ties

[2] The reservation for members of scheduled castes (SC) is 13%, while it is 7% for scheduled tribes (ST), 11% for nomadic tribes and 19% for members of Other Backward Castes (OBC). Within each of these categories, and within the residual "open" category, there is a 30% reservation for female students.
[3] In government and government aided colleges, 100% of the seats available in the college are determined through CAP. Private unaided colleges, on the other hand, have independent control over 15% of the seats, while 65% are determined through CAP.
[4] This cut-off is reduced to 45% for members of scheduled castes.

in the examination score (out of a total of 200) by considering in order: the student's score in Mathematics; in Physics; aggregate marks in the 10th standard examinations; score in Mathematics in the 10th standard examination; and the student's date of birth.

2.2 The effect of the quota system on student composition

The system has the potential to significantly affect the composition of the student body because it does not impose college-specific minimum standard for admission, even in the most selective colleges. Each program *must* admit all students who apply from reserved castes until caste quotas are filled, regardless of the reduction in academic standards this may entail. If students from low castes achieve a grade high enough to warrant admission in the "open" or unreserved category, their numbers are not counted against the quota. Consequently, *all* students admitted into a reserved category seat are those who would not otherwise have obtained admission to the college.

The admission of a relatively large number of under-qualified students into selective programs will necessarily lower the mean ability of students in the program and simultaneously increase variance. The magnitude of this change is determined by two factors: the size of the quota and existing levels of schooling inequality across students who apply for reserved and unreserved seats. In India, the change in student composition as a consequence of the affirmative action policy is large, partly because of the relatively large size of the quota, but primarily because of the significant difference in abilities between reserved and unreserved categories of student.

Data on test results for the 14,696 students who took the 2007 Maharashtra engineering entrance examination (CET) reveal the striking difference in initial schooling ability or academic levels of students from different castes. 84% of students in the top decile (by merit rank) of the ability distribution come from upper castes (UCs) with only *2%* drawn from scheduled castes and tribes (figure 1). In contrast, scheduled caste students (SCs) account for 41% of those in the lowest decile of the ability distribution, while other backward castes (OBCs) account for 42% and upper castes for only 17%.

To get a sense of how the policy affects the composition of students in different programs, figure 2 groups all of Maharashtra's engineering programs[5] into 10 groups ordered by the CET rank cut-off for upper caste male students. For each of these groups, the figure first displays a box plot of the cut-off scores for male upper caste students for all colleges in the

[5] Each field of study offered by a college is considered as a distinct program. The data are by fields of study.

group and then, immediately adjoining it, for male scheduled caste students admitted under the reservation policy. The figure strikingly reveals the extent to which the reservation policy reduces the mean level of (initial) student ability in all colleges and increases its variance. This is particularly true in the most selective colleges. In the lowest quality colleges, the CET rank cut-offs for SC students overlap with that of upper caste students. However, as college quality improves, the maximum cut-off score for SC students in any college in that group falls considerably short of the minimum cut-off score for upper caste students. This is as expected, given the very low numbers of SC/ST students in the top decile of the ability distribution, and their concentration in the lower deciles of this distribution.

2.3 Admission procedures in Maharashtra's Autonomous Colleges

In most Maharashtra colleges, an on-line procedure is used to determine admissions, with students selecting, in order of merit rank, a program out of those in which seats for their reservation category are available. This study, however, is based on data for one of the 5 "autonomous" colleges in the state, which follow an independent admission process. These colleges make their admission decisions prior to the date by which admissions to other Maharashtra colleges must be determined. Students admitted into one of the autonomous colleges are not eligible to participate in the Centralized Allotment Process (CAP) for other engineering colleges in the state (Government of Maharashtra 2007). Because of this, and since the college I consider is by far the highest ranked from amongst the set of autonomous colleges (and amongst the top ranked colleges in the state), students who apply to this college are generally considering admission based on the offerings of this college alone.[6]

On the "counseling" day, all students who are interested in applying for admission to any one of the autonomous colleges must appear in person at a stipulated site. They are then called up by order of their merit rank in the MAH-CET and informed of their choices. Because the process is conducted through a face-to-face interview, significantly more information is provided to the student than would be the case in an anonymous on-line matching system. Instead of just listing their preferences, students can also ask the councilor about the programs they are eligible for and the number of seats which remain, for the student's category, in each program. This provides information on the relative selectivity of the program, given that relative quality, as measured by the difference in the mean CET score of enrolled students across programs, varies over time. This occurs both because several fields are close substitutes for each other but also because, as discussed later in this paper, a significant number of students who are eligible for high quality programs chose less selective ones.

[6] Admission outcomes for the all-India colleges, including the Indian Institutes of Technology, are conducted under a separate process, and outcomes are known prior to students making decisions on state level colleges.

The variation in the difference in mean CET scores across the 9 fields of study in the college is documented to in table 1, which provides data on the mean CET score of students by program and cohort. Though the relative rankings of the programs remains stable, the *difference* in mean CET scores changes over time. Thus, the difference between the mean quality of students in Electronics and Telecommunications relative to computers is substantial for the 2005 cohort, but becomes insignificant for later cohorts. Conversely, the difference between Information Technology and Mechanical is insignificant for the 2005 cohort, but increase over time.

The state-level system of quotas applies to autonomous colleges, with eligibility varying with a student's caste-and gender-specific "category." As a consequence, despite this being one of the state's most selective colleges, there is considerable variation in the quality of students within each program. This primarily reflects variation in ability between castes. Table 2 documents the variation in caste-specific entrance cut-off scores for male students in the 2007 cohort, for each of the 9 fields of study offered by the college. The table confirms the significant difference in the academic ability of students, distinguished by caste, within any given program. This is further attested to in figure 3, which graphs the variation in students' CET rank across the fields for the 2007 cohort, distinguishing between upper caste and scheduled caste students within each field. The figure clearly reveals the variation in initial academic levels within each program. It also reveals that this variation is primarily a consequence of the variation in learning levels across castes, and hence is a consequence of the affirmative action policy.

Critical to the analysis of this paper is the fact that students from the 2005 to 2007 cohorts were organized into first year courses by the field of study that they were admitted to, with one exception. The exception comes from students enrolled in Instrumentation, a field which admits only 30 students. Because of this small class size, the 30 students in this field are divided across the remaining 8 classrooms for the first year program. The empirical analysis of this paper is thus based on the 8 classrooms into which students are grouped for their first year courses, with each classroom comprising primarily of students admitted for the same field of study. [7] The analysis of first year GPA is based only on data for the 2005-2007 cohorts of students, a sample of 1617 students.

The organization of first year classrooms by the fields for which students were enrolled implies that each program of study *can* essentially be treated as a separate "college," with students grouped into classes which differ by the level of course selectivity. This is not true of all engineering

[7] For each student, I have data on the classroom (division) in which they were placed for first year courses, in addition to enrolled field.

colleges or of the 2008 cohort in this particular college. Recognizing the difference in quality across class groupings that this system entails, the college moved to a system which assigns students to first year divisions by an alphabetical ranking of their last names. In such a system, the program a student is admitted into has little implication for academic achievement in the first year, even though it continues to matter from the second year on.[8]

While enrollment choices proceed in order of merit, students do not divide into programs based only on merit; preferences for different fields play a role. This is suggested by the overlap in cut-off scores across fields in figure 3. Stronger evidence of the role of student choice in placements is provided in table 3. For the college which is the subject of this study, the table summarizes the proportion of students in each classroom or division that were admitted into other fields of study (again, each classroom comprises primarily students enrolled in the same discipline, so that organizing the table by enrolled field rather than classroom produces completely similar results). Programs (and hence classrooms or divisions) are ordered by their average selectivity as measured by cut-off scores for male upper caste students in the 2007 cohort.[9] A student is considered eligible for a program if, when he or she is called up by order of rank, the number of category seats taken is less than the total available.[10]

As expected, the table shows that students admitted into any given program were also eligible for less selective programs (minor variations occur because the ranking of programs by selectivity varies across castes and across years). It also shows, however, that a significant proportion of students admitted into any given program were also eligible for admission into programs ranked *higher* in terms of selectivity. For example, as many as 52% of upper caste students enrolled in field 4 were also eligible for field 3. This percentage is 39% for Other Backward Castes, 43% for Scheduled castes and 63% for scheduled and other tribes.

2.4 Academic programs in autonomous colleges

The set of autonomous colleges are distinguished from other engineering colleges in Maharashtra in that they have independent control over their academic program, including the determination of their curriculum and academic standards. Each of these colleges independently determines, for example, the academic criteria each student must satisfy to remain in a program. They also independently determine grading systems

[8] The near random allocation of students to divisions from 2008 on does not lend itself well to analyses of the effects of peer composition on learning, because it generates little variation in either the mean or variance of ability across divisions.
[9] The ranking of programs by selectivity varies slightly by caste. So, for example, for some castes, and in some years, program 2 is considered to be more selective than program 1.
[10] Calculating eligibility this way is equivalent to seeing if the student's CET score, within their caste and gender category, exceeded the score of the last admitted student.

in each course, the division of students into classes for first year programs, and the overall curriculum each student must follow.

All students follow a common course of study in the first year, focusing on Mathematics, Physics, Chemistry and introductory classes in basic engineering fields (Engineering Mechanics, Computer Programming, Basic Electronics, Engineering Graphics, Electrical Engineering and Tech Shop). The common subject matter in the first year is particularly conducive to the analysis of this paper, since all students take the same examinations and are ranked on a common standard. Thus, it is possible to compare the cumulative grade point average (GPA) of students, regardless of the program they are enrolled in. It also implies that even though students are divided into classrooms by their chosen field, there is no variation in their access to laboratories or other resources, including teacher quality. All classrooms are taught by a set of teachers who specialize in the core subjects.

The straightforward comparison of the GPA of students across different divisions is also enabled by the "relative grading" policy adopted by the college, in which faculty members jointly agree on a common grading system, further ensuring that grades can be compared across sections and courses. This is in contrast to the system in U.S. colleges, for example, where each professor may adopt their own grading policy, making difficult comparisons of grades achieved under different professors. For the college in question, a uniform set of tests is administered in each subject, with agreement by faculty on how each question will be graded. Grade point averages are assigned on a 10 point scale, offering more variation than is the case in most U.S. colleges. The college also has a well-defined criterion for failing a student: A student fails if he or she receives a GPA of less than 5 in three consecutive semesters.

2.5 Data and Regression Sample

For the college I study, I have data for the 2005-2008 student cohorts on their cumulative grade point average in each year, their merit rank and score in the CET examination, whether they were admitted in a reserved seat and, if so, the type of seat (eg. scheduled caste male or female), their caste and gender, field of study and the division (classroom) in which they were placed in the first year.

I use data only from the 2005-2007 cohorts, because it is only for these cohorts that first year classes were organized on the basis of the fields in which they were enrolled. The analysis is further confined to students from upper castes, other backward castes, scheduled castes and to males from scheduled and nomadic tribes.

Women from scheduled and nomadic tribes are omitted because quotas for this group, in each subject, are rarely filled, so that women are essentially eligible for all programs. This is clearly revealed in table 4, which specifies the percentage of students in each caste- and-gender category, which are eligible for each of the 8 fields of study. Women belonging to scheduled and nomadic tribes are eligible for almost all fields of study. The resulting sample size is 1,536 students spread over three cohorts.

3. Theoretical Framework

The empirical analysis of this paper builds on a theoretical model in which individuals choose a field of study and a college, based on their skill set and the difference in income which different fields earn. For simplicity in notation, and without any loss of generality, I assume each field of study within a college constitutes a distinct program, indexing programs by j=1,…J. I therefore ignore an independent role for college quality, whereby students may choose a lower ranked program if it is offered in a highly regarded college.[11]

I assume that each individual i is endowed with a vector of skills, $\mathbf{s}_i$, which takes the form of a set of competencies which are differentially valued in each field and in the general program of studies undertaken in the first year. The field-specific vector of valuation prices (or weights) is represented by $\boldsymbol{\tau}_j$, j=1,…J. For example, some fields may require high computational skills, while others may be more dependent on analytical skills. Thus, a student may do exceptionally well in one particular field of study, to which he or she is particularly well suited, but not in others. This assumption is similar to that in the Roy model of self-selection across occupations, in which the market generates differential returns to the different skills with which individuals are endowed (Roy 1950, Heckman and Sedlacek 1985). An individual's skill set is unobserved by the econometrician but known to the individual.

In addition to an individual's endowment of skills, performance also depends on the quality or selectivity of the program, represented by a vector of variables, $\overline{A}_j$. In the empirical analysis, $\overline{A}_j$ is represented by the mean and variance of the (initial) ability of students in a program. $\overline{A}_j$ influences learning not just through classroom peer effects, but also because highly selective programs may attract better teachers and have more resources. Finally, learning is also affected by a set of individual demographic characteristics, such as gender and caste, X, as well as the student's initial ability level, A^o.

[11] Since the empirical analysis is based on the programs offered by a single college, any college specific returns to learning are absorbed in the regression constant.

I index students by their rank (i) and category (c), so that student i of category c has one higher rank than student i+1 of the same category. For each cohort, enrollment decisions are made in the year prior to admission, so that the cohort that enters college in year t makes its decisions in year t-1, based on information available on each program at year (t-1), I_{t-1}. For simplicity in notation, in this section I consider only outcomes for the cohort which enters college at year t, making decisions in year (t-1), and so do not include year subscripts unless necessary.

At time t-1, expected academic achievement of student i of category c in program j is based on his or her expectation of student composition in the program in period t:

$$(1)\quad E(A_{icj} \mid I_{t-1}) = \alpha_1 A_{ic}^{o} + \alpha_2 E(\overline{A}_j \mid I_{i,t-1}) + \alpha_3 X_i + s_i{}' \tau_j + \upsilon_i^{a} \quad j = 1,\ldots J$$

The income an individual earns on graduation from college depends on his or her observable academic achievement (A_{icj}), and on a set of individual demographic characteristics (X_i). The market additionally generates a return to attending selective programs, with earnings increasing with $\overline{A}_j$. The same skills (**s**) which result in students having comparative advantage in different fields of study are also differentially rewarded in the market place, with the market return to the skill set being denoted by the vector π. Thus, income (Y) for an individual i of category c who graduated from program j is[12]:

$$(2)\quad Y_{icj} = \gamma_1 A_{icj} + \gamma_2 \overline{A}_j + \gamma_3 X_i + s_i{}'\pi + \upsilon_i^{y} \qquad j = 1,\ldots J$$

I define the indicator variable $d_{icj}=1$ if student i of category c is eligible for program j (the discussion of eligibility rules follows later in this section). Let S_i be the set of all programs j such that $d_{icj}=1$. From amongst this set, the student will chose the program which yields the highest expected income, based on his or her (t-1) information set. That is, the student will choose $j \in S_i$, if:

$$(3)\quad E((Y_{icj} - Y_{ick}) \mid I_{t-1}) \geq 0 \qquad j, k \in S_i,$$

[12] This specification assumes that the set of skills, S, is observed in the market place. This would follow if the model is re-interpreted as one of life time earnings, with skills being observed over time. An alternative model would allow employers to use an individual's major as indicative of his or her skill set, generating an income function in which returns vary across majors. The difference between these two sets of assumptions does not alter the empirical analysis of this paper.

Let δ_{icj} be an indicator variable, taking the value 1 if student i of category c *enrolls* in program j, 0 otherwise. Substituting (1) and (2) into (3), a student enrolls in program j ($\delta_{icj}=1$) if the student is eligible for the program and if:

$$(4)\quad [\beta_1 \mathrm{E}\,((\bar{A}_j - \bar{A}_k) \mid I_{i,t-1}) + s_i'(\tau_j - \tau_k)] \;\geq\; 0 \qquad \forall j,k \in \mathrm{S_i}$$

Turning to eligibility, a student is eligible for program j ($d_{icj}=1$) if the number of higher-ranked students of the same caste- and gender-category who have *accepted* the program is less than the number of category seats available. The total number of students of higher category rank who have accepted the program, D_{icj}, the summation of δ_{mcj}, m=1 to i-1, depends on the skill set of all previous students ($\mathbf{S}_{i-1}$) and the category rank of student i. Let N_{cj} be the total number of seats available for students of category c in program j. Then eligibility for a program is given by the following condition:

$$(5)\qquad d_{icj}(S_{i-1}, N_{cj}, i) = 1 \quad if \quad D_{icj}(\mathrm{S_{i-1}}, i) = \sum_{m=1}^{i-1} \delta_{mcj} < N_{cj}$$

$$0 \quad \text{otherwise}$$

If the system operated as a pure meritocracy, eligibility for a program would depend only on a student's merit rank relative to available seats. With student choice, eligibility also reflects the field-specific abilities of all students of higher rank.

As noted above, enrollment in a program requires both choice (eq. 4) and eligibility (eq. 5) constraints to be satisfied. Students who choose program j include those for whom it is the most selective program they gain admission to, but also those who get into more selective programs but choose the less-selective one because they are particularly suited to it.

The importance of an individual's skill endowment in determining his or her placement outcomes will vary with the selectivity of the program. Consider the most selective program. Of the students who meet the cut-off requirements for this program, those who enroll in it will be those for whom the returns to their skill set is also higher in this program, but also those for whom the returns in other programs are high enough, but not sufficiently so as to overcome the difference in program quality. In the least selective programs, however, the set of students enrolled will include this set, but also those who got admitted into better quality programs, but the returns to the lower quality program far exceeded that of the higher quality program. That is, lower quality programs are more likely to include students who have selected into the program on the basis of their skills, and hence those who are likely to do better in the program than predicted either by the initial ability level or by the quality of the program.

4. Empirical Framework

I start this section by specifying the source of bias in estimates of academic achievement. I then describe the empirical methodology of this paper, and subsequently turn to a discussion of identification. The section concludes by describing the methodology used to allow for heterogeneity in estimates of the effect of student composition on learning.

4.1 Bias in estimates of academic achievement

If students did not differ in skills, the distribution of students across programs would solely reflect eligibility criteria (5), and would therefore be a non-linear function of their category rank. Conditioning on this non-linear function or, alternatively, exploiting a strong regression discontinuity framework which compares outcomes for students close together in category rank but who differ in the programs in which they were placed, would generate unbiased estimates of the effect of student composition on learning. Alternatively, if program-specific skills were observable, unbiasedness could also be obtained by conditioning regressions on the skill set.

The empirical challenge in estimating the effect of program selectivity $(\bar{A})$ lies in the fact that students' program-specific skills are likely to determine enrollment choices and are unobservable. Equation (4) suggests that some students who are admitted into a higher quality program, $\bar{A}_{jk}, \bar{A}_{jk} \geq \bar{A}_{lm}$, may nevertheless chose a field which is ranked lower because they are endowed with program-specific skills which enable them to do better in these programs than might otherwise be expected. The data discussed in section 2.3 confirms that this is often the case. Consequently, the returns to mean student ability will be biased downwards, unless selection bias is accounted for in the empirical analysis.

Specifically, from equation (1), and with skill endowments being unobservable, a student's academic achievement is given by:

$$\text{(6)}\quad \mathrm{E}(A_{icj} \mid A^0_{ci}, X_i, \delta_{ic}) = \alpha_1 A^o_{ic} + \alpha_2 E(\bar{A} \mid \delta_{ic}) + \alpha_3 X_i + \mathrm{E}(u_i \mid A^0_{ic}, X_i, \delta_{ic})$$

where $u_i = s_i' \tau + \upsilon_i^a$. Because the vector $\boldsymbol{\delta}_{\mathbf{ic}}$, representing the student's choice amongst J programs, affects the returns to his or her skill set, the regression error term is not independent of $\boldsymbol{\delta}_{\mathbf{ic}}$. $\boldsymbol{\delta}_{\mathbf{ic}}$ also determines the value of $\overline{A}$, the measure of college quality which directly affects learning. As a consequence, OLS estimation of (6) will generate biased estimates of α_2, the effect of college selectivity on learning.

This bias exists even when students enrolled in different fields pursue a common course of study in core subjects, as they typically must do in the first year of most engineering colleges. Since academic achievement in any field is likely to depend on core competencies, but with the extent of this dependence varying across fields, a student's choice of field of study is also informative about his endowment of the skills needed for core subjects. Put differently, knowledge of **δ,** the indicator vector of placement outcomes, provides information on both the quality of the program the student is enrolled in as well as his endowment of core skills. And, as before, since the extent of selection by ability is likely to be greater in less-selective programs, ignoring selection is likely to generate a downward bias in estimates of the effect of selectivity on academic achievement.

4.2 *Empirical methodology*

As previously noted, because enrollment is *not* determined exclusively by a student's merit rank, identification on the basis of a strong regression discontinuity design is not possible. An alternative instrumental variable approach to the identification of the effect of $\overline{A}$ is rendered difficult because of a lack of suitable instruments. While determinants of eligibility rules appear to be natural instruments for $\overline{A}$, the regression error term also varies with enrollment outcomes, δ, and hence with the determinants of eligibility. This invalidates the use of eligibility cut-offs, for example, as instruments in a conventional instrumental variables regression.

Identification is achieved by using $E(u_i \mid A_i^o, X_i, \delta_i)$ as a control function to purge the correlation between the regression error term and δ (Heckman and Robb 1985, 1986; Florens, Heckman, Meghir and Vytlacil 2008). Under assumptions regarding the joint distribution of u_i and η_{i},, the vector of control functions, $\mathrm{K}_i = E(u_i \mid A_i^o, X_i, \delta_i)$ can be estimated and included amongst the regressors in the estimation of (7). This generates the new regression equation:

$$(7)\ \mathrm{E}(A_{ic} \mid A_{ic}^0, X_i, \delta_{ic}) = \alpha_1 A_{ic}^o + \alpha_2 E(\overline{A} \mid \delta_{ic}) + \alpha_3 X_i + \mathrm{K}_i + E(u_i - \mathrm{K}_i \mid \delta_{ic}, A_{ic}^o, X_i)$$

Identification follows if the following conditions are met:

(8a) K_i depends on δ_{ic}

(8b) $E(u_i - K_i \mid \delta_{ic}, A_{ic}^o, X_i) = 0$

Since the conditional mean $E(u_i \mid A_{ic}^o, X_i, \delta_{ic})$ is a function of Pr(δ_{icj} =1 | Z_{icj}), including control functions amongst the regressors essentially conditions on the individual's enrollment outcomes. Identification assumes that, conditioning on this choice, the qualities or attributes of the program that the student is enrolled in are orthogonal to the regression error term. A student chooses his field of specialization. Once he does so, the attributes of the program are given. Because the correlation between $\overline{A}_j$ and U_{ij} arose only as a consequence of **δ,** assumption (8b) implies that $(U_{ij} - K_{ij})$ is independent of $\overline{A}_j$, in regressions which condition on K_{ij}.

Rewriting (7), the regression equation I estimate takes the following form:

$$(9)\, E(A_i \mid A_i^0, X_i, \delta_i) = \alpha_1 A_i^o + \alpha_2 \overline{A}_j + \alpha_3 X_i + \sum_{j=1}^{J} \alpha_{4j} \delta_{ij} E(u_{i|} \mid \delta_{ij} = 1, A_i^0, X_i)$$

As noted by Heckman and Robb (1985, 1986), because the control function $E(u_i \mid A_{ic}^o, X_i, \delta_{ic})$ is a function of Pr(δ_{icj} =1|Z_{ij}), it can be approximated by a polynomial function of this probability. That is, rather than make an assumption regarding the bivariate distribution of (U_i, η_i), it is possible to make a weaker assumption regarding the distribution of η_i. I assume that this is normally distributed, and so approximate $E(u_i \mid A_{ic}^o, X_i, \delta_{ic})$ by $\Phi(Z'\beta)$ and $(\Phi(Z'\beta))^2$.

In turn, $\Phi(Z'\beta)$ can be estimated by a simple probit estimation of enrollment outcomes (reflected in enrollment indicators, δ_{icj}). As discussed in the previous section, enrollment in any program reflects eligibility as well as students' acceptance of the admission offer. It is always possible to represent the multiple decision rules (4) and (5) which determine a student's enrollment in program j, j=1,..J, by a single index function IN:

$$(10)\, IN_{icj} = f(Z_{icj})'\beta + \eta_i, \quad Z_{icj} = \{\overline{A}_j - \overline{A}_k, s_{ic}, S_{i-1}, N_{cj}, i\} \;\; \forall\; j,k \in S_i,\; j \neq k$$

With δ_{icj}=1 if $IN_{icj} = f(Z_{icj})'\gamma + \eta_i >= 0$; 0 otherwise. Estimation of (10) provides estimates of $\Phi(Z'\beta)$.

Since the regression equation includes estimates of the control function, it is necessary to correct standard errors for the use of estimated variables. The calculation of standard errors is, however, difficult because the regression includes J different control functions, one for each field. All regressions therefore use block bootstrapped standard errors, which allow for correlation across error terms for students of the same cohort who are enrolled in the same program.

4.3 Identification

Estimation of (9) requires an exclusion restriction: a variable in Z which is not in X. Since the vector K is a function of the probability that the student is enrolled in each of the J available programs, determinants of eligibility for each program appear to constitute valid instruments since eligibility is typically determined by a set of school-defined rules, on the basis of variables which may be uncorrelated with unobserved student ability.

However, when eligibility is determined by a cut-off entrance exam score, the use of an eligibility indicator for identification of the control function is problematic for several reasons. First, as is well recognized, it identifies enrollment outcomes only for those observations which fall in the neighborhood of the cut-off score. Second, because identification exploits a specific discontinuity in the relationship between initial test scores and subsequent learning, it is important to allow for a (direct) non-linear relationship between initial test scores and learning outcomes. Doing so often reduces the precision of estimates.

The setting of this paper reduces these concerns to some extent. As in Bertrand et al's (2010) study, identification is aided by the fact that India's higher educational system combines features of a meritocracy with an affirmative action program. Consequently, eligibility for any given program is determined by a student's *category* rank, their rank within their reservation caste-gender category, not their overall merit rank. Further, I consider outcomes in 8 different programs, with eligibility criterion being defined separately by gender and for 8 different caste groups.[13] Each program has a total of 60 seats, divided amongst 16 caste-gender groups. Because of the relatively large number of admission categories, identification is not just at one margin and regression estimates are therefore far more precise. While it is unlikely that a student's *category* rank affects his or her performance, any direct role of category rank on learning can be allowed for by including a quadratic in the individual's category rank amongst the regressors.

[13] The large number of caste groups with separate quotas is a feature of Maharashtra's reservation system. Bertrand et al (2010) study outcomes in a different state in India, with reservations applying to relatively few caste groups.

Equation (10) also suggests, however, that the choice to enroll in any program j will reflect students' expectations of a vector of contrasts between the quality of students in that program and in other programs. Information on the variables in students' (t-1) information set, which determine their time t expectations of quality, can thus be a source of identification, if they do not directly influence outcomes.

The richness of the data set and the unique nature of the admission process provide information on students' time (t-1) information set, and hence on their expectations of student composition in different programs. As previously described, enrollment decisions are made sequentially, by order of merit, so that when any individual student is called upon to make his enrollment decision, students of higher rank have already made theirs. Each student therefore knows the number of seats, of the category total, available in each program, or, equivalently, his or her class category ranking – their ranking within their reservation category for each program. This determines eligibility for any given program, but also represents the only information available to the student at the time of the enrollment decision on the (expected) composition of students in each division and hence of the selectivity of different programs. Programs in which the student is ranked lower can be expected to be of higher quality, since they have already enrolled more students with higher test scores. A comparison of program category rank across different programs thus provides information on ex post differences in the quality of the programs the student is eligible for. Because I have data on the merit rank and enrollments of all students in the college, it is possible to construct the very same information that is available to students at the time of their enrollment decision.

Specifically, for each student, program j is considered to be more selective than j+1 if the number of students of higher rank that have committed to program j exceeds the number that have committed to program j+1. Since the highest rank a student can achieve is 1, the numerical value of a student's (category) rank will be greater in a more selective program than in a less selective one. As long as program selectivity affects choice, the probability of choosing program j+1 will fall with the difference in category rank between programs j and j+1.[14]

For each student eligible for program j, I construct two variables, $mdiff_{icj}$, the difference in the student's class category rank between the most selective program he or she is admitted to and program j (with selectivity

[14] The difference in the student's *overall* class rank in two fields may also matter for admission choices if students care about where they rank in the class' overall distribution. However, a comparison of the student's category rank does not confer this same information. Thus, students of upper classes are likely to be at the top of the distribution in all programs, while students of scheduled castes and tribes are likely to be at the bottom, regardless of their category rank.

defined as above), and $ldiff_{icj}$, the difference in class category rank between program j and the next most selective program he or she is admitted to. These two variables, and their square, in addition to the eligibility indicator, constitute instrumental variables which determine program placement in regressions on schooling achievement (GPA). Both $mdiff_{icj}$ and $ldiff_{ic}$, vary across students and, for each student, across programs.

Since each student compares not just the most selective program and program j, but also other programs, it is necessary to also condition enrollment outcomes on the student's overall category rank and rank in the program in question. However, these rankings are not used for identification since they may reflect ability. I allow them to directly affect learning by including them in the regressor set used in the second-stage estimation of learning equations.

$Mdiff_{icj}$ and $ldiff_{icj}$ satisfy the necessary criterion for valid instruments. First, they are likely to influence enrollment outcomes because they represent the only information that students have, at the time of making their enrollment choices, on class composition and the quality of their future peers. Data on student composition for previous cohorts, or indeed on cut-off scores, is not publicly available,[15] so that they cannot be a component of student's time (t-1) information set. Students do, of course, have beliefs regarding the relative ranking of programs and are likely to use these to inform their enrollment choices. However, because many of the programs are close substitutes for each other and because class composition varies from year to year (as evidenced in table 1), these beliefs are generally formed at a fairly aggregate level. For example, students generally state that Electronics and Telecommunication, Computers, and Information Technology are amongst the top programs without being able to definitively rank one over the other. Similarly, Production and Metallurgy are clubbed together at the bottom of the rankings. This set of aggregate beliefs does not provide information on how the quality of students in Information Technology, in any given year, will contrast with those in Computers, and hence does not help students choose between these closely related fields. Conversely, information on how peers are selecting between the two programs provides exactly this information. The significance of relative class category rank in enrollment choices is, of course, a testable hypothesis; first stage regression results are presented in the next section.

A vector of contrasts based on student's class category rank in different programs is also likely to satisfy the second criterion for identification: It is unlikely to directly influence outcomes, based as it is on partial information which varies across students. This partial information is

[15] The Government of Maharashtra has started providing data on cut-off scores and on students enrolled in different programs, from the 2007 cohort on. However, data from the autonomous colleges are not included.

unlikely to affect classroom learning, which will depend on ex post measures of student quality, based on the (completed) admission outcomes of all students. These completed measures do not vary within a classroom, and hence are unlikely to be correlated with the individual-specific measures of program selectivity which inform enrollment decisions.

In this, it improves significantly over the use of a vector of contrasts based on ex post (completed) realizations of classroom composition. Both difference vectors use information on other programs, and it can be argued that this serves to identify enrollment in any given program. However, it is likely that the assignment of other school inputs, including teachers, may also reflect the relative ex post quality of different programs. In contrast, the placement of teachers cannot be based on the individual-specific information available to each student at the time of their enrollment on the enrollment outcomes of higher-ranked students.

4.4 Heterogeneity

An additional benefit of the control function estimator is that it is robust to heterogeneity in the effects of student composition on learning (Heckman and Robb 1985, 1986). This is particularly important in this context, because the large variance in ability within a classroom implies that students will correspondingly differ, quire significantly, in the extent to which their ability levels differ from those of the mean student. Correspondingly, the effect of student composition is also likely to differ across students.

To allow for heterogeneity, I measure the difference between each student's entrance test score and the mean for the classroom *(dist)*, and interact this variable with the mean and variance of classroom ability. I further allow the effect of the interacted term to differ for students who are above the mean *(TOP)* and those who are below it.

This specification is preferred to a more conventional treatment for heterogeneity, which allows estimates to vary across students at different quantiles of the ability distribution, primarily because it is better suited for comparing outcomes across alternative admission regimes which differ in the distribution of ability within a program. Methods which categorize students by their place in the ability distribution generate estimates which are specific to a distribution; any estimated effect of being in the top quartile of the ability distribution need not transfer over to a regime which generates little variation in ability across students, thereby blurring the difference across students in the top quartile relative to others. In contrast, the methodology of this paper is robust to changes in the distribution; it allows for reduced heterogeneity in estimates in regimes characterized by significant lower levels of ability variance.

5. Summary Statistics and Simple Regression Evidence on Caste-based differences in academic achievement

Table 5 provides data on the merit rank, CET scores and first year CGPAs of students of 3 different cohorts, who initiated their studies in 2005, 2006 and 2007 respectively. The data are reported separately for members of upper castes, other backward castes (OBCs), scheduled castes (SCs) and scheduled and other tribes. The data reveal the striking difference in initial academic achievement, even across students in one college, particularly between members of scheduled castes and tribes, relative to other students. For the 2007 cohort, for example, the mean rank of students from upper castes was 1086, while it was 4395 for scheduled castes and as low as 6856 for scheduled and other tribes. This difference in academic ability persists after one year in the engineering program: The CGPA of upper caste students significantly exceeds that of students from Other Backward Castes, Scheduled Castes and Scheduled Tribes. F tests for the significance in CGPA across upper caste students and students from reserved caste, reported in the table, reveal the difference to be statistically significant in all cases, except for the difference between upper caste students and OBCs for the 2005 cohort.

OLS regressions allow us to probe some of the factors underlying this difference in achievement across castes, and provide preliminary evidence on the effects of classroom composition on learning. I start with a simple regression, which regresses the CGPA of the 2005-2007 cohorts of students on indicator variables for OBCs, SCs, and STs. To maintain comparability with later regression results, this regression sample omits female members of scheduled and other tribes. The data on sample means reported in the previous table suggests the importance of allowing for cohort effects. Consequently, the regression also contains dummy variables for (2005 and 2006) cohorts. The variation exploited in the paper is thus variation within cohorts. The regression also includes a dummy variable for gender, and interactions of gender with caste categories. All standard errors in this and subsequent regressions are clustered by cohort, division and caste.

The results from this simple regression are reported in the first column of table 6. They confirm the very significant difference in achievement across students of different castes, with students from scheduled tribes faring most poorly, followed by students from scheduled castes. Controlling for cohort effects, the difference between the average CGPA of OBCs and upper castes is 0.36, while it is as much as one point lower for SCs and 1.4 points lower for STs. The difference is significant in all cases.

The second column in the table assesses the extent to which this difference reflects students' initial ability level, with regression (1) being augmented to include the student's score in the CET entrance examination as well as his or her overall merit and category rank. In this and all subsequent regressions, rank varies inversely with initial achievement with the top student receiving a rank of 1. A positive effect of initial achievement, as measured by rank, therefore implies a negative coefficient on the *Merit rank* variable. The regression results reveal that differences in initial achievement levels explain much of the unconditional effect of caste on a student's GPA; conditioning on initial achievement levels, the magnitude of the coefficients on caste indicators is significantly reduced. The difference in GPA across upper caste students and OBCs and SCs is 0.3, while the difference in CGPA across upper castes and scheduled tribes is 0.4.

The final regression reported in this table includes the mean entrance exam (CET) score and the variance in this score within the classroom. OLS estimates suggest a positive, but statistically insignificant effect of mean initial achievement levels of students, and a negative effect of the variance. The effect of variance is statistically significant at the 5% level.

6. Results

6.1 First Stage Enrollment Regressions

Table 7 reports estimates from probit regressions of the probability that an individual enrolls in a program. Regressors include the *eligibility* indicator as well as measures of each student's assessment of the quality of the program relative to the most selective program he or she was admitted to (*mdiff)* as well as of the next most selective (*ldiff)*, as well as their squared terms. To ensure that the coefficient on these terms do not merely reflect a non-linear relationship between enrollment and a student's category rank or category rank within the program in question, the set of regressors also includes the student's overall and class category rank and their square. Other included variables are the set of individual characteristics used in the OLS regression of the previous table (the student's entrance examination score and rank, dummy variables for caste and gender, and interactions of these indicators) as well as a set of cohort dummy variables. Not surprisingly, these regressions confirm that the indicator variable for eligibility significantly affects enrollment outcomes.

As expected, eligibility has a strong positive effect on enrollment in any program. However, the results also confirm that students' enrollment choices reflect the information available to them, at the time of admission, on other programs. Thus, in almost all cases (enrollment in program 1 being the exception), an increase in the student's assessment of the relative selectivity of the most selective program he or she is admitted to, relative to the program under consideration, reduces the probability of enrollment in that program, while an increase in the selectivity of the program being considered, relative to the next most selective program the student is eligible for increases the probability of enrollment. In the vast majority of cases, the effect of these relative selectivity indicators is statistically significant at the 5% level.[16]

6.2 The effect of the mean and variance of student ability on first year CGPA

Table 8 reports regression results of the determinants of students' first year CGPA. The first regression reports results from an OLS regression which differs from those previously reported in table 6 in the inclusion of a student's class category rank and rank squared, as well as the square of the category rank, amongst the regressors. In addition to these variables, the set of regressors include the mean and variance of students in the classroom, and all other variables used in the OLS regressions of table 6.

The second column reports results using control functions for each of the 8 programs derived from the probit estimates reported in table 7. An F test for the joint significance of the 16 control function terms (based on the probability of eligibility and its square for each of the 8 programs), reported at the bottom of this column, confirms the importance of selectivity bias.

Both OLS and control function regressions generate insignificant effects of the mean ability level of students in a classroom on individual learning, though the control function estimate exceeds in magnitude that from OLS regressions. In contrast, the effect of variance in mean ability is negative and significant in all specifications, with the control function estimates being approximately twice the magnitude of OLS estimates. The coefficient estimate of -0.002 translates into an elasticity of -0.86, evaluated at the mean of the relevant variables. This suggests that a 10% increase in the variance of initial ability levels in the classroom reduces individual CGPAs by 9%.

6.3 Testing heterogeneity in the effects of classroom composition

As discussed in Section 4, I test for heterogeneity by incorporating interactions of the mean and variance of initial ability with the difference between the student's test score and that of the mean student, further allowing the effects to vary across students with scores above and below that of the mean student.

[16] The fact that these variables play no significant role in explaining enrollment in the most selective program suggests that selection is less likely to be an issue here, with enrollment more likely to reflect merit.

The first regression in table 9 reports the results from this specification. The results suggest that the effect of mean student ability as well as those of variance are similar for students at the top and bottom of the distribution (F tests, reported at the bottom of the table, fail to reject the hypothesis of equal effects). Interactions with the difference between student ability and mean ability generate coefficients for mean ability which are opposite in sign and statistically significant for high ability students, but similar in sign and statistically insignificant for variance (F tests for equality are again reported at the bottom of the table).

Regression 2 implements the coefficient restrictions suggested by the first regression, additionally including an interaction of mean student ability with the square of *dist.* The regression suggests that students at the top and the bottom of the ability distribution are most adversely affected by the affirmative action policy. High ability students are adversely affected by the reduction in mean ability generated by the policy, while low ability students are adversely affected by being placed in high quality programs.

While the magnitude of the effect of the variance in student ability on learning remains the same as in previous specifications without heterogeneity (coefficient of -0.002), the effect of mean ability increases to 0.10. In standard deviation units, a 1 standard deviation increase in the mean ability of students raises CGPA by 1.4, while a 1 standard deviation increase in the variance increases CGPA by 0.2.[17]

6.4 Regression Results for second year CGPA

Table 10 reports regression results for second year CGPA, using the same control functions from the previous set of regressions which allow for heterogeneity in the effect of student composition on individual learning. This regression allows us to examine whether the effects of first year classroom composition persist over time. I restrict the analysis only to second year CGPA, because of relatively high drop-out rates in the third and fourth year of the program. In contrast, there are few drop-outs from the program in the first two years, partly a consequence of a college policy which does not fail students before the completion of two years of study. The overall drop-out rate after 1 year is 5%.[18] By caste, first year drop-out rates are 3% for upper castes and Other Backward Castes, 5% for Scheduled Castes and 8% for tribal castes. At the end of two years, the drop-out rate is 8%, varying from 6% for upper castes and OBCs, 8% for

---

[17] A one standard deviation change moves a student from the middle of the distribution (the fiftieth percentile) to the 84^{th} percentile, with effects being the largest for students at the center of the distribution (on account of the normality assumption).

[18] The regression sample is restricted to students who complete the second year of studies. Since this is approximately 95% of all students, any bias introduced by sample selection is likely to be insignificant.

SCs, and 15% for tribal castes. Drop-out rates increase significantly in the third year, to 12% for the sample as a whole (10% excluding tribal castes). By caste, the third year drop-out rate is 9% for upper castes and OBCs, 11% for SCs and as high as 24% for tribal castes.

Results from regressions based on students second year CGPA are qualified by the fact that second year courses differ by field, so that not all students take the same set of courses. This is in contrast to the first year course of study, in which all students, regardless of field, take the same courses. Consequently, in contrast to the analysis of second year CGPA, concerns regarding differences in the degree of difficulty of the subject matter across fields do not arise. However, even in the second year program, the college's relative grading system, which strives to make students' CGPA comparative across fields, helps mitigate these concerns to some extent.

As before, I report results from two specifications which are identical to those reported in the previous table (table 9) for first year CGPA. The first regression allows for interactions of the mean and variance of student ability with the difference between the student's test score and the mean, distinguishing between students with scores which exceed the mean and those with scores which are lower. The second regression implements some of the exclusion restrictions suggested by the first, allowing the student's place in the ability distribution to affect the coefficient on mean ability, but not variance.

The regression results suggest an even larger effect of mean student ability on second year CGPA, and a smaller, though still statistically significant effect of variance. Heterogeneity persists, with students at the top and the bottom of the ability distribution being most affected by the affirmative action policy. Thus, student composition continues to affect learning, even in the second year of the program.

7. Predicting academic performance under alternative admission criteria

The estimates of the previous section can be used to infer how students would fare under alternative admission rules, in those cases where it is possible to predict which program each student would be placed in, and hence the mean and variance of ability within each program as well as each student's place within the program. In what follows, I provide estimates of first year CGPA, by caste, under three different admission criteria.

First, as a benchmark case, I consider what outcomes would be under a pure meritocracy, in which students are assigned to programs solely on the basis of their rank in the entrance examination, allowing neither for any affirmative action by caste or for the self-selection of students into lower quality programs. For this exercise, I use the information on the rank of the 14,700 students who took the engineering entrance examination in 2007, and order all students by their rank. Because most programs have an enrollment of 60 (equal to the pupil teacher ratio stipulated by the All India Council of Technical Education), I divide this ordered list of students into blocks of 60, and calculate the mean and variance of student achievement for each of these groups. Matching students to group by their CET rank provides an approximation of classroom composition (measured again in terms of the mean and variance of the CET score) for each student, were admissions to be determined purely on merit.

I then estimate achievement levels under the assumption that each program stipulates a minimum qualifying level for reserved castes, enrolling those who are do not meet this standard but who would normally qualify (under a quota system without this minimum requirement) into a preparation course which brings them up to the necessary level. This admission policy is relevant, since it mimics the system currently in place in India's premier engineering colleges, the Indian Institutes of Technology.[19] The results of this exercise can be interpreted as indicative of the achievement levels that would result if it were possible to improve initial ability levels of all reserved category students.

I consider two different qualifying levels. The less restrictive policy imposes a qualifying level equal to the current cut-off score for scheduled caste students, thereby effectively requiring all scheduled tribe students, 18% of the students in each program, to achieve this level prior to admission. The more restrictive policy sets the qualifying score at the minimum score achieved by an OBC student. This affects 31% of students in each program. In each case, for students who fail to meet this new cut-off, I reassign their score to equal this cut-off (under the assumption that, on completion of the remedial course, they have to be at the minimum qualifying standard to gain admission). I then re-calculate the mean and variance of student ability in each program. Individual CET scores are, however, maintained at their true level, so that, for the purposes of this prediction exercise, the only change is in the mean and variance of achievement levels in each program.

[19] Under this system, reserved category students are only admitted if their entrance examination score exceeds a stipulated level, set at two-thirds of the lowest score recorded by an open category student. To ensure compliance with quotas, rejected reserved category students who would have been admitted without this minimum cut-off level are enrolled in a one-year remedial education program in Mathematics, Physics, Chemistry and English. On completion of the program, students in this class are required to take an examination to ensure that they meet the minimum entrance requirement for the field. If so, they are admitted into the first year of the regular program.

Table 11 describes the characteristics of the student population of the programs in the surveyed college under the current system, the meritocracy, and the two minimum standards system (SC cutoff and OBC cutoff). The current affirmative action policy of ensuring that reserved category students have equal access to the most selective programs yields a system in which there is no significant difference in the quality of the programs that SC students are placed in, relative to OBC students and students from upper castes. This is true also of the two minimum standards systems, since they also ensure reserved castes access to all programs.

In contrast, in a pure meritocracy, upper caste students and those from OBCs are placed in far more selective programs. The mean CET score of students in the programs in which upper caste students are placed, for example, improves from 168 to 180. Conversely, SC students experience a very large decline in program quality, with mean CET scores falling from 168 to 156. The decline for ST students is even more severe, from a mean score of 168 to 145.

The move from the current system to a meritocracy also implies, for all castes, a huge reduction in the variance of ability in classrooms in which they are placed, with variance falling to close to zero. While variance still remains high in the minimum standards system, it is substantially reduced relative to the current system. The estimates in table 11 suggest that variance in the OBC-cut off system is approximately 4 standard deviations lower than that in the current system. Thus, the reduction in variance substantially exceeds the improvement in mean ability, which amounts to approximately 0.6 standard deviations for upper caste students.

Particularly for the pure meritocracy, but also for the minimum standards regime, the reduction in variance also implies that heterogeneity will be reduced, because of the reduction in the difference in individual test scores from the classroom mean. Thus, the improvement in the mean student ability for upper castes under a meritocracy is likely to increase CGPA for all students of this caste, while the corresponding decrease in student quality is likely to reduce learning for reserved category students. However, because the reduction in variance is so significant, the effects of changes in mean ability are likely to be dominated by the effects of reduced variance. This is likely to be particularly true for the two minimum standards regimes, which generate only small changes in the mean ability of students.

Predicted CGPA under the current system and the three alternative admission criteria are in table 12. All students do better under a meritocracy though, as expected, the gains are significantly larger for students from upper castes and from OBCs, suggesting the significant cost to these students of the current quota system. For upper caste students, mean CGPA improves from 7.5 to 8.1, an improvement of approximately one-half of a standard deviation in each year. While this may seem small in magnitude, the few studies which assess the effect of academic achievement on income suggest significant returns, even to small improvements in CGPA. Available studies are mostly based on data from the U.S., and examine the effect of high school GPA in earnings. Hanushek and Woessmann (2008), reviewing this literature, report an average effect of a 12% increase in annual earnings in response to a one standard deviation improvement in test scores. They also report data from developing countries which suggests that the return to achievement in these countries may be even higher. While no estimates exist of the returns to academic achievement in Indian engineering colleges, the fact that, in this system, each student's CGPA and even ranking within a program is a matter of public record, and hence known to employers, suggests that the returns to improvements in CGPA are likely to be high.

The striking feature of table 12 is that even the less restrictive SC cut-off generates significant improvements in learning (relative to the current system), yielding estimates of CGPA close to those in a meritocracy. The more restrictive OBC cut-off generates CGPA estimates which essentially equal those obtained under a pure meritocracy for upper castes, OBCs and SCs, and exceed it for scheduled tribes. This reflects the fact that this system significantly improves the mean level of students in programs in which ST students are placed, even while reducing variance relative to the current system. Raising the performance of SC and ST students to that of the weakest OBC student thus generates significant returns.

These results suggest that quota systems need not imply a reduction in quality relative to a pure meritocracy; it is possible to do away with the extreme caste-based segregation that a pure meritocracy would imply and still maintain learning levels. Thus, the adverse effect of the current system on learning (relative to a meritocracy) essentially reflects the very low (current) ability levels of students from scheduled castes and tribes. An effective affirmative action policy would therefore require combining a system of quotas in higher education with policies addressed at improving learning levels for students from backward castes at the elementary and secondary levels.

The results which maintain under a meritocracy are qualified by the very large reduction in variance, relative to the current system, which a meritocracy entails. Because the lack of variance under the meritocracy places it substantially out of the range of variances currently observed, it is an open question whether the same regression coefficients would apply in an environment of limited variance.

Further, this paper focuses on one dimension of program quality, student composition. It is unlikely that the estimates include the effect of other schooling inputs, such as teacher quality. This is because identification is based on students' expectations of the difference in the composition of students across programs within one (highly ranked) college. Within this college, teacher quality and the availability of other schooling inputs do not vary significantly, so that the estimates likely identify a "pure" effect of student composition. Conversely, the large reduction in program quality which members of reserved castes would face under a meritocracy would also probably imply reductions in the quality of teachers and other schooling resources. Thus, the estimates of learning under a pure meritocracy are best viewed as upper bound estimates for students from scheduled castes and tribes.

These concerns do not, however, arise in the minimum standards case, both because the reduction in variance is far less acute and also because such a change in admission criteria would not affect the colleges in which students are placed and hence resource quality.

8. Conclusion

Despite the considerable debate on the merits of affirmative action programs, there are few studies which provide evidence on its benefits and costs, particularly for the quota systems which exist in developing countries such as India. The effects on learning have been difficult to substantiate because of the endogeneity problems which must be solved in order to credibly estimate how students are affected by the changes in the mean and variance of the ability distribution of students in any program as a consequence of the affirmative action policy.

This paper uses the meritocratic design of the Indian higher education system for identification, exploiting a specific feature of the admission process whereby students must make enrollment choices based on an information set which provides incomplete information on program selectivity and ultimate student composition in each program. Using this information to construct a set of control functions, I control for the college placement outcomes of each student. Conditional on college placement, the attributes of the college are orthogonal to the regression error term.

At the onset, it is important to note that the results of this paper are based on data for only one college, one of Maharashtra's most selective colleges. The extent to which they would generalize to data on students from the middle of the ability distribution of all students is an open question. However, this being said, the results are informative on how the most selective colleges are affected by the quota system. Because the quality of education at the best colleges generates externalities which likely affect the entire labor force, these results are of significant interest. Additionally, the descriptive evidence on the mean and variance of student ability by caste in all the engineering colleges in the state (figure 2), suggest that it is the top end of colleges which are most affected by the affirmative action policy.

The regression results suggest that the affirmative action policies reduce learning for students of all castes. This primarily reflects the large discrepancy in ability between upper caste and scheduled caste students at the onset of higher education. As a consequence, the policy of ensuring that all programs, regardless of their degree of selectivity, provide admission to the same fixed percentage of scheduled caste and tribe students, without regard to their level of learning, results in programs characterized by substantial variance in ability levels. I find that all students are hurt by the variance in ability within a program that the current system generates. Students at the top and bottom end of the academic distribution are also hurt by the changes in mean student ability which the policy entails.

To infer the costs of the system, I compared results from the current system to a benchmark case of a pure meritocracy, the system which would result if the affirmative action program were removed and students were placed in colleges solely on the basis of their scores in entrance examinations. This comparison suggests that the policy generates a lower level of cognitive skills. This result is, however, qualified by the large difference between the variance in ability in the current system and in a meritocracy; predictions so far out of the sample range may not be accurate.

An alternative scenario in which the initial ability level of reserved castes is increased to a stipulated minimum standard also reduces the costs of the current system. This experiment suggests that learning in higher education institutions could be improved by policy efforts to raise the ability levels of reserved castes at the primary and secondary levels of schooling, rather than attempting to do so at the higher education level. In the absence of earlier corrective measures, the large difference in the ability levels across scheduled castes and tribes and upper caste students generates persistently lower learning at the higher education level for all students.

Introducing affirmative action policies at an earlier stage would also help address a major criticism of systems which focus on correcting inequalities at the higher education level: it redresses schooling inequalities only for those among the scheduled castes and tribes who complete high school with a grade that makes them eligible for admission to college. In urban Maharashtra, data from the Government of India's National Sample Surveys reveal that only 27% of the SC/ST students between the ages of 28 and 25 in 2004 completed grade 12, compared with 43% from upper castes. Data on per capita expenditure of those SC/ST households whose children have completed grade 12 reveal that they are relatively wealth compared to other households from this caste. Their monthly per capita expenditure is

Rs. 6,613, while those of households whose children do not complete 12th grade is only Rs. 3,622.

The results of this study thus suggest that the most effective affirmative action policy would be one which addresses caste-based schooling inequalities at the primary and secondary level. Though the current system may generate an income gain to scheduled castes (Bertrand et al 2010), the empirical evidence linking cognitive skills to income, both at the individual level and at the level of the macro-economy (Hanushek and Woessmann 2008) suggests that these income gains would be higher under alternative policies which still ensure reservations for scheduled castes but reduce the gap in ability by caste at the onset of higher education. Such a policy would benefit not just reserved caste students, but also students of higher castes.

References

Arcidiacono, Peter. 2005. "Affirmative Action in Higher Education: How do Admission and Financial Aid Rules Affect Future Earnings?"

Bertrand, Marianne, Rema Hanna and Sendhil Mullainathan. 2010. "Affirmative Action in Education: Evidence from Engineering College Admissions in India." *Journal of Public Economics* 94:16-29.

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." *The Quarterly Journal of Economics*, November: 1491-1527.

Ding, Weili and Steven F. Lehrer. 2007. "Do Peers Affect Student Achievement in China's Secondary Schools." *The Review of Economics and Statistics* 89(2): 300-312.

Duflo, Esther, Pascaline Dupas and Michael Kremer. 2008. "Peer Effects and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." Cambridge, Mass: National Bureau of Economic Research, working paper 14475 (available at http://www.nber.org/papers/w14475).

Florens, J.P., J. J. Heckman, C. Meghir and E. Vytlacil. 2008. "Identification of Treatment Effects using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects." *Econometrica* 76(5):1191-1206.

Government of Maharashtra. Directorate of Technical Education. 2007. *Rules and Institute Information Brochure for Academic Year 2007-08.* Mumbai: Government of Maharashtra.

Hanushek, Eric A., John F. Kain, Jacob M. Markman and Steven G. Rivkin. 2003. "Does Peer Ability Affect Student Achievement?" *Journal of Applied Econometrics* 18:527-544.

Hanushek, Eric A., John F. Kain and Steven G. Rivkin. 2002. "New Evidence about Brown v. Board of Education: The Complex Effects of School Racial Composition on Achievement." Cambridge, MA: National Bureau of Economic Research, working paper 8741 (available at: http://www.nber.org/papers/w8741)

Hanushek, Eric A. and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46(3):607-668.

Heckman, James J. and Richard Robb, Jr., 1985. "Alternative Methods for Evaluating the Impact of Interventions." In James J. Heckman and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data.* New York: Cambridge University Press.

Heckman, James J. and Richard Robb. 1986. "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on outcomes." In Howard Wainer (Ed.). *Drawing Inferences from Self-selected Samples* (Lawrence Erlbaum Associates).

Hoekstra, Mark. 2009. "The Effect of Attending the Flagship State University on Earnings: A Discontinuity-based Approach." *The Review of Economics and Statistics* 91(4):717-724.

Holzer, Harry and David Neumark. 2000. "Assessing Affirmative Action." *Journal of Economic Literature,* 38(3): 483-568.

Hoxby, Caroline Minter. 2000. "Peer Effects in the Classroom: Learning from Gender and Race Variation." Cambridge, MA: National Bureau of Economic Research. Working paper no 7867 (available at http://www.nber.org/papers/w7867)

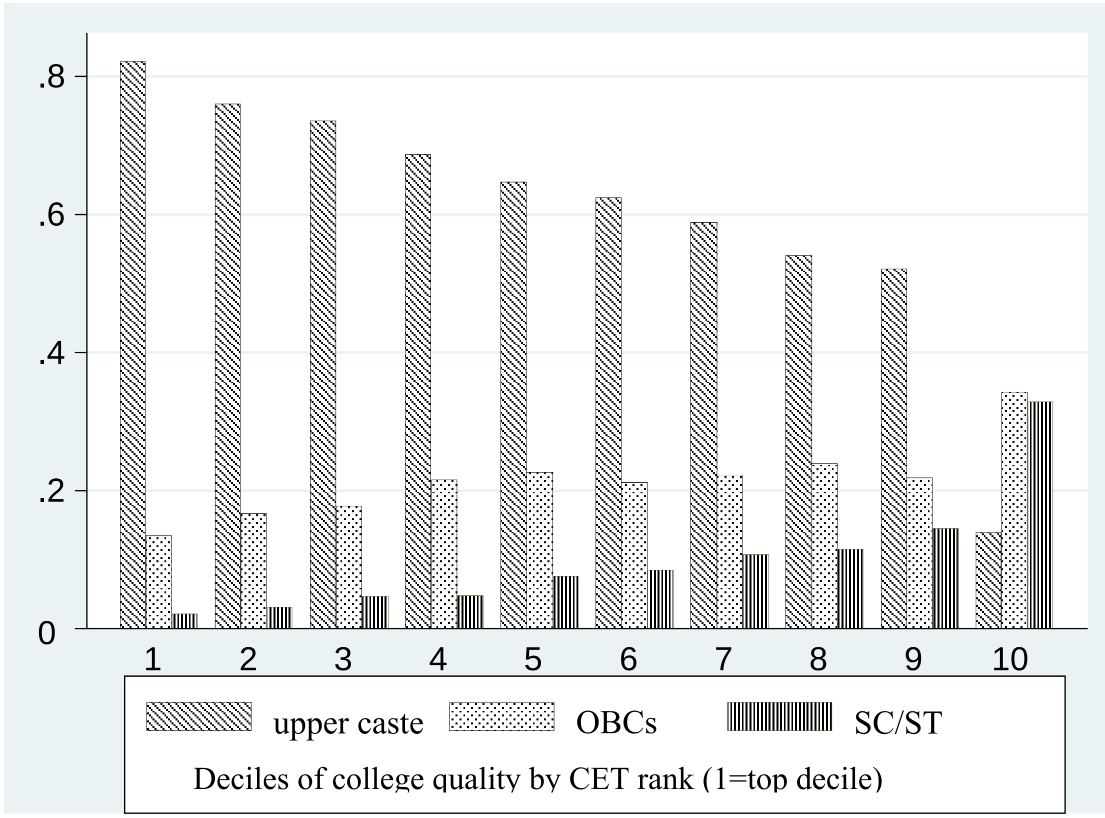

Figure 1: Caste-wise distribution of first year students, all Maharashtra engineering colleges 2007-08, by decile of college quality

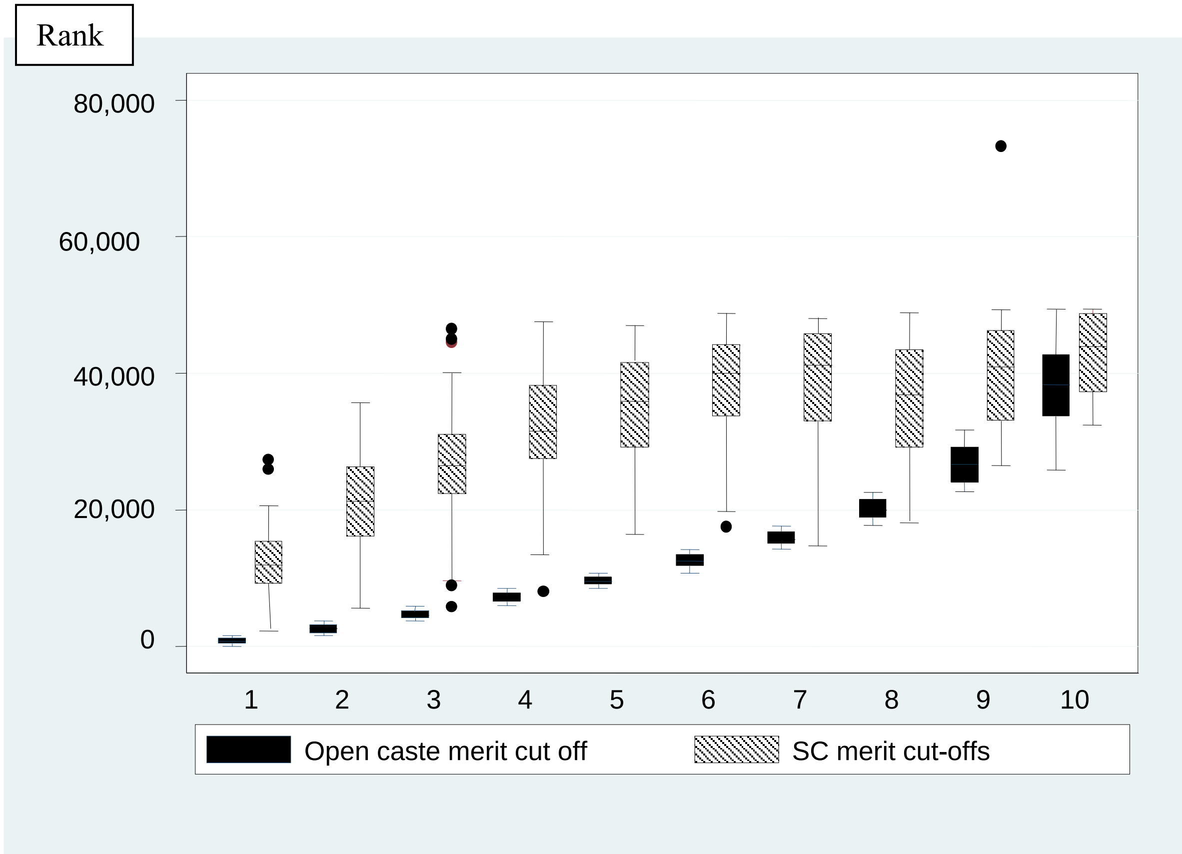

Figure 2: Merit rank cut-off scores, open category students and scheduled castes, by deciles of college quality, all Engineering Colleges, Maharashtra, 2006.

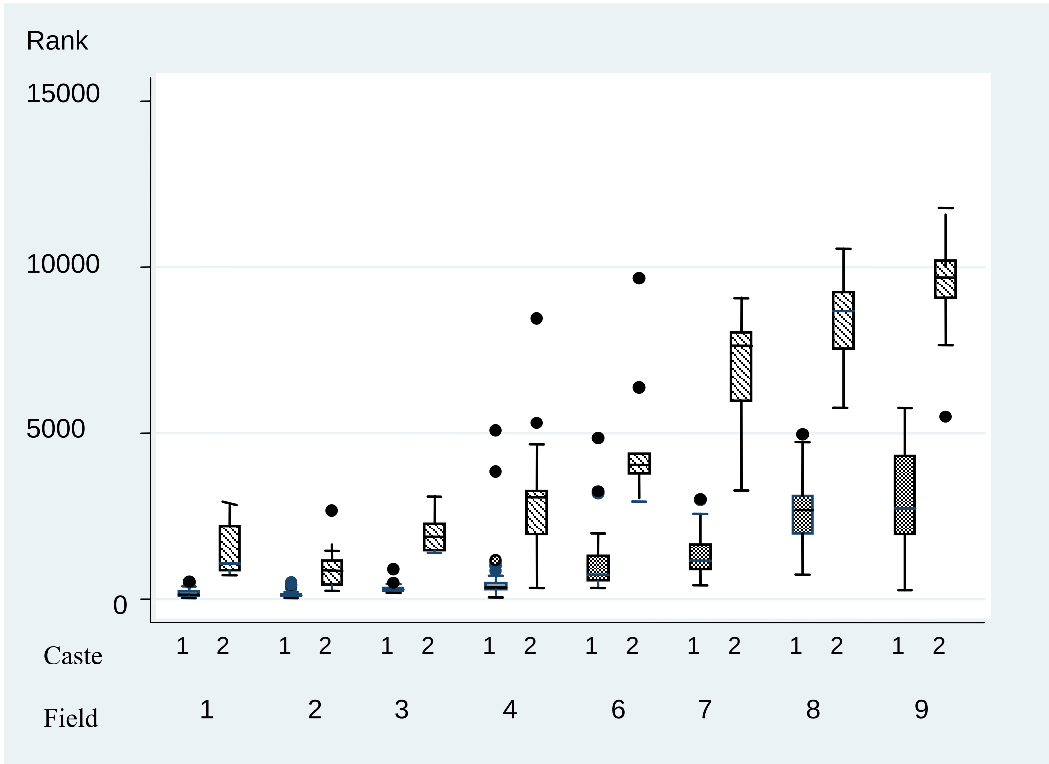

Figure 3 :Spread of rank by subjects in survey college, 2007 cohort, upper caste (1) and scheduled caste (2) students

Table 1.: Mean CET scores of students, by field and cohort

Field	2005 cohort	2006 cohort	2007 cohort
Electronics and telecommunication	176.2	183.9	183.9
Computers	169.8	185.0	183.4
Information Technology	165.5	176.7	179.5
Mechanical	164.4	172.2	172.9
Instrumentation	159.8	173.3	172.1
Electrical	156.3	168.6	166.8
Production	147.8	162.9	161.3
Civil	138.6	154.4	152.2
Metallurgy	132.0	155.7	148.5

Source: Survey data

Table 2: Cut-off Entrance Examination scores for male students by caste, 2007-08, in survey college

	Open	OBC	SC	ST
Computer	191	186	177	130
Electronics and Telecomm	190	184	178	148
Information Technology	188	182	168	124
Mechanical	186	177	161	106
Instrumentation	184	174	163	104
Electrical	178	173	155	100
Production	174	168	137	136
Civil	163	155	126	109
Metallurgy	161	151	123	128

Source: survey data

Table 3: Eligibility of students for different programs (ranked by selectivity), by enrolled program and caste

Enrolled Program	Proportion of students who are eligible for program							
	1	2	3	4	5	6	7	8
Prog. 1								
Open	0.99	0.90	0.99	0.99	1.0	1.0	1.0	1.0
OBC	0.97	0.72	0.95	1.0	1.0	1.0	1.0	1.0
SC	0.96	0.91	0.96	1.0	1.0	1.0	1.0	1.0
ST	0.94	0.70	0.97	0.91	0.97	1.0	1.0	1.0
Prog. 2								
Open	0.87	0.97	0.96	1.0	1.0	1.0	1.0	1.0
OBC	0.78	1.0	1.0	1.0	1.0	1.0	1.0	1.0
SC	0.62	1.0	1.0	1.0	1.0	1.0	1.0	1.0
ST	0.55	1.0	0.97	0.94	0.97	1.0	1.0	1.0
Prog. 3								
Open	0.05	0.23	0.97	0.99	1.0	1.0	1.0	1.0
OBC	0.25	0.06	1.0	0.94	1.0	1.0	1.0	1.0
SC	0.04	0.17	0.96	0.96	1.0	1.0	1.0	1.0
ST	0.21	0.29	0.93	0.79	0.93	1.0	0.96	1.0
Prog. 4								
Open	0.12	0.23	0.52	0.97	0.99	0.99	1.0	1.0
OBC	0.07	0.06	0.39	1.0	0.87	1.0	0.97	1.0
SC	0.14	0.24	0.43	0.90	0.86	1.0	1.0	0.94
ST	0.32	0.33	0.63	0.98	0.90	1.0	0.87	0.98
Prog. 5								
Open	0.03	0.04	0.07	0.22	0.91	0.96	0.99	1.0
OBC	0.05	0.03	0.05	0.38	0.88	0.9	0.93	1.0
SC	0.0	0.04	0.04	0.13	0.88	0.96	0.96	1.0
ST	0.15	0.23	0.23	0.38	0.92	0.96	0.85	0.96
Prog. 6								
Open	0.01	0.01	0.01	0.02	0.27	0.97	1.0	1.0
OBC	0.0	0.0	0.0	0.12	0.02	0.27	1.0	0.86
SC	0.08	0.08	0.08	0.08	0.2	0.96	0.88	0.92
ST	0.27	0.23	0.35	0.31	0.38	0.96	0.85	0.81
Prog. 7								
Open	0.0	0.0	0.02	0.04	0.11	0.32	0.93	0.96
OBC	0.0	0.0	0.0	0.12	0.02	0.27	1.0	0.88
SC	0.05	0.05	0.05	0.05	0.19	0.57	0.95	0.86
ST	0.1	0.1	0.1	0.45	0.4	0.65	0.95	0.8
Prog. 8								
Open	0.02	0.03	0.04	0.04	0.07	0.24	0.69	0.97
OBC	0.12	0.09	0.16	0.23	0.16	0.33	0.60	0.93
SC	0.0	0.0	0.0	0.05	0.05	0.5	0.63	1.0
ST	0.33	0.24	0.33	0.38	0.38	0.57	0.66	1.0

Source: Survey data

Table 4: Percentage of students eligible for different programs, by gender and caste, 2007 cohort

Caste	Program							
	1	2	3	4	5	6	7	8
Men								
Open	24	46	45	57	79	88	98	100
OBC	29	19	38	54	68	77	91	98
SC	29	31	39	61	69	81	92	100
ST	32	46	43	89	100	100	86	100
TC-A	25	31	38	94	56	100	81	86
TC-B	20	30	40	50	70	80	90	100
TC-C	19	13	50	44	63	75	81	94
TC-D	60	20	40	50	70	100	100	90
Women								
Open	19	16	31	60	63	81	93	100
OBC	5	15	26	100	72	100	95	100
SC	26	15	30	44	52	100	96	100
ST	33	50	67	100	100	100	100	100
TC-A	33	100	100	100	100	100	100	100
TC-B	100	20	100	100	100	100	100	100
TC-C	100	100	20	100	100	100	100	100
TC-D	100	100	100	100	100	100	100	100

Source: survey data

Note: TC-A, TC-B, TC-C and TC-D refer to different nomadic tribes, each with their own reservation quota.

Table 5 : Average merit rank and CET score of students in survey college, by cohort and caste

	2005 cohort	2006 cohort	2007 cohort
Merit Rank			
Upper castes	2090.19	1227.40	1086.15
	(1963.16)	(1242.41)	(1301.56)
Other backward	2492.67	1800.79	1893.60
castes	(1737.62)	(1289.96)	(1673.14)
Scheduled castes	4551.96	4444.78	4394.93
	(2123.88)	(3302.04)	(3302.40)
Scheduled &	4234.04	6024.63	6855.6
nomadic tribes	(2290.80)	(5135.55)	(4693.14)
CET score			
Upper castes	163.83	179.07	180.33
	(18.64)	(10.92)	(11.42)
Other backward	158.63	173.36	172.88
castes	(16.07)	(10.38)	(12.62)
Scheduled castes	140.15	158.01	157.11
	(16.88)	(17.46)	(19.38)
Scheduled &	143.23	150.74	144.66
nomadic tribes	(19.32)	(23.86)	(25.60)
First year CGPA			
Upper castes	6.86	7.27	7.57
	(0.97)	(1.25)	(1.24)
Other backward	6.70	6.84	7.09
castes	(0.87)	(1.07)	(1.13)
Scheduled castes	5.89	6.32	6.36
	(1.04)	(1.09)	(1.32)
Scheduled &	5.25	4.98	5.26
nomadic tribes	(1.32)	(1.89)	(1.52)
F tests for equality			
Upper caste	2.19	11.67	14.04
CGPA=OBC	(0.14)	(0.001)	(0.00)
Upper caste	45.98	37.11	52.70
CGPA=SC	(0.00)	(0.00)	(0.00)
Upper caste	21.72	57.92	76.61
CGPA=ST	(0.00)	(0.00)	(0.00)

Note: Standard deviations in parentheses. For F tests, the figures in parentheses report the probability of the test statistic exceeding the estimated value.

Table 6. OLS regression estimates of the effect of caste on first year Cumulative Grade Point Average

	Regression 1	Regression 2	Regression 3
CET score	--	0.03*	0.03*
		(0.01)	(0.01)
Merit Rank	--	-0.005	-0.004
		(0.004)	(0.004)
Category rank		-0.001	-0.001
		(0.001)	(0.001)
Mean CET, class	--	--	0.001
			(0.01)
Variance CET, class	--	--	-0.001*
			(0.0004)
Other backward castes (OBC)	-0.36*	-0.27*	-0.26*
	(0.16)	(0.10)	(0.09)
Scheduled castes (SC)	-1.01*	-0.28*	-0.24+
	(0.17)	(0.13)	(0.13)
Scheduled tribes (ST)	-1.38*	-0.42	-0.37*
	(0.23)	(0.25)	(0.16)
Female	0.29*	0.28*	0.28*
	(0.10)	(0.08)	(0.08)
OBC*female	0.03	0.24+	0.25*
	(0.16)	(0.13)	(0.13)
SC*female	-0.16	0.05	0.07
	(0.20)	(0.16)	(0.16)
Sample size	1536	1536	1536
Regression F (prob. > F)	12.04 (0.00)	67.14 (0.00)	80.3 (0.00)

Note: Regression is run on first year data for three cohorts of students. The sample includes students from upper castes, other backward castes, schedule castes and males from scheduled and nomadic tribes, omitting nomadic tribes and scheduled tribe females. In addition to variables listed above, the regressors include dummy variables for different cohorts and for gender. Figures in parentheses are standard errors, corrected for clustering by cohort, division and caste.

* Significant at 5% level

\+ Significant at 10% level

Table 7: Probit Estimates of the Probability of Enrollment in different programs

	Enrollment in program							
	1	2	3	4	5	6	7	8
Eligibility	1.69*	1.92*	1.05*	1.83*	1.34*	1.36*	1.11*	0.43
	(0.26)	(0.25)	(0.22)	(0.24)	(0.22)	(0.27)	(0.29)	(0.32)
ldiff	-0.02	0.08	0.21^{+}	0.26*	0.13	0.22*	0.14*	0.15*
	(0.08)	(0.07)	(0.11)	(0.06)	(0.09)	(0.09)	(0.06)	(0.07)
ldiff square	0.001	-0.003	-0.02*	-0.01*	0.001	-0.0003	-0.005	-0.003
	(0.004)	(0.003)	(0.007)	(0.003)	(0.005)	(0.005)	(0.004)	(0.004)
mdiff	0.03	-0.13*	-0.18*	-0.15*	-0.24*	-0.16*	-0.11*	-0.13*
	(0.05)	(0.05)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)
mdiff square	-0.001	0.005^{+}	0.01*	0.006*	0.01*	0.004*	0.004*	0.003*
	(0.002)	(0.003)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
Category	0.01	0.02*	0.02*	-0.004	0.02*	0.02*	0.01*	0.02*
Rank	(0.01)	(0.01)	(0.01)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
Category	-0.0001	-0.0001*	-0.0002*	0.00	-0.0001*	-0.0001*	-0.0001*	-0.00005*
rank sq	(0.00004)	(0.00)	(0.00005)	(0.00)	(0.00)	(0.00003)	(0.00002)	(0.00002)
Class category	0.13*	0.07	0.12	0.02	0.07	0.004	0.04	-0.07
rank	(0.07)	(0.05)	(0.08)	(0.03)	(0.07)	(0.08)	(0.04)	(0.07)
Class category	-0.008*	-0.006*	-0.005	-0.0003	-0.01*	-0.004	-0.003^{+}	0.001
rank square	(0.003)	(0.002)	(0.003)	(0.0005)	(0.003)	(0.003)	(0.002)	(0.004)
CET score	0.03^{+}	0.002	-0.03*	-0.03*	0.02^{+}	-0.07*	-0.06*	-0.06*
	(0.01)	(0.01)	(0.017)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Merit rank	0.006	-0.003	-0.03*	-0.01*	-0.01	-0.03*	-0.01	-0.02
	(0.01)	(0.01)	(0.01)	(0.006)	(0.01)	(0.01)	(0.01)	(0.007)
Sample Size	1536	1536	1536	1536	1536	1536	1536	1536
LR χ^2 (13)	421.29	407.81	334.34	563.52	332.23	383.14	363.04	392.16
Prob > χ^2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Notes: Additional regressors include gender, caste and interactions, and cohort fixed effects. Regression sample is as in previous tables.

*Significant at 5% level $^{+}$Significant at 10% level

Table 8: OLS and Control Function Estimates of the Effect of Student Composition on First Year Schooling Achievement (CGPA)

	OLS	Control Function
Mean CET	-0.001	-0.01
	(0.006)	(0.008)
Variance CET	-0.001+	-0.002*
	(0.0005)	(0.001)
Category Rank	-0.004	-0.007*
	(0.003)	(0.003)
Category rank square	0.00002	0.00003*
	(0.00)	(0.00001)
Class category rank	0.004	0.01
	(0.009)	(0.01)
Class category rank square	-0.0002	-0.002
	(0.0002)	(0.002)
CET	0.03*	0.03*
	(0.008)	(0.01)
Merit rank	-0.004	-0.004
	(0.004)	(0.005)
OBC	-0.27*	-0.31*
	(0.09)	(0.12)
SC	-0.28*	-0.33*
	(0.13)	(0.15)
ST	-0.41*	-0.43*
	(0.16)	(0.15)
Female	0.27*	0.23*
	(0.08)	(0.08)
OBC*female	0.25*	0.26*
	(0.13)	(0.13)
SC*female	0.06	0.14
	(0.16)	(0.16)
Sample Size	1536	1536
Wald χ^2 (32)	F(16,111)=72.77	1511.21
(Prob > χ^2)	(0.00)	(0.00)
F test for joint significance of control functions	--	28.16
(prob >F)		(0.03)

Notes: All regressions include cohort fixed effects. Standard errors (in parentheses) are clustered by cohort, division and caste. Control function estimates use bootstrapped standard errors.
*Significant at 5% level
+Significant at 10% level

Table 9: Testing for heterogeneous effects of student composition on First Year Schooling Achievement (CGPA) - Control Function Estimates

	(1)		(2)	
	Coeff	Std. Error	Coeff	Std. error
Mean CET	--	--	0.10^{*}	(0.05)
Variance CET	--	--	-0.002^{*}	(0.001)
Top*Mean CET	0.084^{*}	(0.043)	--	--
(1-Top)*Mean CET	0.085^{*}	(0.043)	--	--
Top* Variance CET	-0.001	(0.0008)	--	--
(1-Top)* variance CET	-0.002	(0.001)	--	--
Top*Mean CET*dist	0.001^{*}	(0.0003)	0.001^{*}	(0.0002)
(1-Top)*Mean CET*dist	-0.0004	(0.0003)	-0.001^{+}	(0.0003)
Top*Variance CET*dist	-0.0001	(0.0001)	--	--
(1-top)*Variance CET*dist	0.00	(0.00)	--	--
Top*Mean CET*dist sq	--	--	2.56 e-6	5.17 e-6
(1-top)*Mean CET*dist sq	--	--	1.19 e-6	1.17 e-6
Sample Size	1536	--	1536	
Wald χ^2 (Prob > χ^2)	1520.11	(0.00)	1337.82	(0.00)
χ^2 test for joint significance of control functions (prob > χ^2)	29.07	(0.02)	16.73	(0.04)
Test for equality of Mean CET (Prob > χ^2)	0.33	(0.56)	--	--
Test for equality of variance (Prob > χ^2)	0.46	(0.50)	--	--
Test for equality of Mean CET*dist (Prob > χ^2)	3.92	(0.04)	4.28	(0.04)
Test for equality of Variance CET*dist (Prob > χ^2)	0.68	(0.41)	--	--

Notes: All regressions include cohort fixed effects, caste and gender dummies and interactions, caste and class caste rank and their squares, student's CET and merit rank. Bootstrapped standard errors (in parentheses) are clustered by cohort, division and caste. *Significant at 5% level $^{+}$Significant at 10% level

Table 10: Effect of student composition on Second Year Schooling Achievement (CGPA) – Control Function Estimates

	(1)		(2)	
	Coeff	Std. Error	Coeff	Std. error
Mean CET	--	--	0.15^{*}	(0.05)
Variance CET	--	--	-0.001^{*}	(0.0006)
Top*Mean CET	0.16^{*}	(0.06)	--	--
(1-Top)*Mean CET	0.16^{*}	(0.06)	--	--
Top* Variance CET	-0.001	(0.001)	--	--
(1-Top)* variance CET	0.00001	(0.001)	--	--
Top*Mean CET*dist	0.001^{*}	(0.0005)	0.001^{*}	(0.0003)
(1-Top)*Mean CET*dist	-0.001^{+}	(0.0004)	-0.001^{+}	(0.0003)
Top*Variance CET*dist	-0.0001	(0.0001)	--	--
(1-top)*Variance CET*dist	-0.00002	(0.0001)	--	--
Top*Mean CET*dist sq	--	--	-1.06 e-6	4.63 e-6
(1-top)*Mean CET*dist sq	--	--	8.43 e-7	1.89 e-6
Sample Size	1481	--	1481	
Wald χ^2 (Prob > χ^2)	835.00	(0.00)	1076.28	(0.00)
χ^2 test for joint significance of control functions (prob > χ^2)	39.60	(0.001)	42.86	(0.0003)
Test for equality of Mean CET (Prob > χ^2)	1.13	(0.28)	--	--
Test for equality of variance (Prob > χ^2)	0.73	(0.39)	--	--
Test for equality of Mean CET*dist (Prob > χ^2)	4.21	(0.04)	7.83	(0.01)
Test for equality of Variance CET*dist (Prob > χ^2)	0.08	(0.77)	--	--

Notes: All regressions include cohort fixed effects, caste and gender dummies and interactions, caste and class caste rank and their squares, student's CET and merit rank. Bootstrapped standard errors (in parentheses) are clustered by cohort, division and caste. *Significant at 5% level +Significant at 10% level

Table 11: Program Composition by caste, in the current system, a pure meritocracy, and a minimum standards system – 2007 cohort

Classroom attributes	Current System	Pure Meritocracy	SC cut-off	OBC cut-off
Upper Castes (n=243)				
Mean CET in classroom	168.37	180.18	169.80	173.56
	(9.89)	(11.72)	(11.82)	(10.29)
Variance CET in classroom	373.28	0.22	231.86	88.65
	(67.77)	(0.38)	(128.25)	(54.93)
Other Backward Castes (n=129)				
Mean CET in classroom	167.64	172.73	169.06	172.96
	(10.14)	(13.02)	(11.86)	(10.36)
Variance CET in classroom	373.33	0.13	234.22	87.55
	(67.87	(0.28)	(122.18)	(53.77)
Scheduled Castes (n=81)				
Mean CET in classroom	168.07	156.46	169.26	173.24
	(9.85)	(19.29)	(11.91)	(10.38)
Variance CET in classroom	380.76	0.08	245.77	91.78
	(55.56)	(0.10)	(130.39)	(55.56)
Scheduled and nomadic tribes (n=90)				
Mean CET in classroom	168.48	143.62	170.01	173.81
	(9.90)	(24.75)	(11.62)	(10.10)
Variance CET in classroom	372.29	0.07	228.27	84.80
	(69.42)	(0.11)	(123.94)	(51.97)

Note: Classroom composition in a pure meritocracy is predicted using data on all 14,700 students who took the engineering examination in 2007, by grouping students, by rank, in groups of 60. SC and OBC cutoff refer to figures obtained by determining eligibility in a program by the cut-offs which apply for these two groups respectively. Figures in parentheses are standard deviations.

Table 12: Predicted CGPA under the current system and in a pure meritocracy, by caste

	Predicted First Year CGPA			
	Current System	Meritocracy	SC cut-off	OBC cut-off
Upper Castes	7.5	8.1	7.8	8.0
	(0.04)	(0.03)	(0.04)	(0.03)
F test for equality with current system (Prob >F)	--	F(1,217)=2802.6 (0.00)	F(1,242)= 406.73 (0.00)	F(1,242)=3322.2 (0.00)
Other Backward Castes	7.1	7.7	7.4	7.6
	(0.06)	(0.06)	(0.06)	(0.05)
F Test for equality With current system (Prob > F)	--	F(1,113)=1808.4 (0.00)	F(1,128)=212.9 (0.00)	F(1,128)=1719.2 (0.00)
Scheduled Castes	6.4	6.8	6.6	6.9
	(0.10)	(0.12)	(0.11)	(0.10)
F test for equality With current system (Prob > F)	--	F(1,74)=260.25 (0.00)	F(1,79)=98.59 (0.00)	F(1,79)=1048.1 (0.00)
Scheduled tribes	5.7	5.9	6.0	6.3
	(0.11)	(0.16)	(0.11)	(0.11)
F test for equality With current system (Prob > F)	--	F(1,78)=13.72 (0.00)	F(1,89)=169.882 (0.00)	F(1,89)=1365.3 (0.00)