

**Analysis of Undergraduate Admissions
to University of California Campuses by Race and Ethnicity
May 2005**

Technical Appendix

What was the research question?

At each UC campus, do similarly qualified applicants from different racial or ethnic groups have the same likelihood of admission?

What methodology was used to answer this question?

1. We divided the applicant pool at each campus into clusters of similar applicants. This was done using two different statistical methods: logistic regression and tree models (also called “recursive partitioning”). Both methods produce estimates of the probability of admission for each individual applicant at the campus under consideration, based on his or her quantitative academic and demographic characteristics. Using each method, we grouped applicants into approximately 20 clusters of students with similar probabilities of admission. (For example, with the logistic regression model, the first cluster contained applicants whose estimated probability of admission was 0-5%, the next cluster was 5-10%, and so on, up to 95-100%.) Race and ethnicity were not factors in grouping applicants into clusters; i.e., students with similar characteristics were assigned to the same cluster regardless of their racial/ethnic group.
2. Within each cluster of similar applicants, we predicted the number who would have been admitted, for each racial/ethnic group, using a procedure that estimates outcomes that would occur under an admissions process that is free of racial or ethnic influence. Under such an admissions process, applicants from different racial/ethnic groups but within the same cluster should be admitted at about the same rate – specifically the overall admit rate for that cluster. Therefore, within each cluster, we predicted the number of admits for a racial/ethnic group by multiplying the overall admit rate for the cluster times the number of applicants from the racial/ethnic group that are within the cluster. For example, if 33% of all the applicants in a cluster were admitted, and if there were 60 Chicano/Latino applicants in that cluster, then the predicted number of Chicano/Latino admits for that cluster would be $33\% \times 60 = 20$.
3. For each racial/ethnic group, we calculated the total number of predicted admits and compared this to the number of actual admits. The total number of predicted admits for a racial/ethnic group is the sum of the predicted number of admits for that group across all clusters. The charts on the left-hand side of Figures 4-11 compare these predicted numbers of admits to the actual numbers of admits for each racial/ethnic group. The charts on the right-hand side of Figures 4-11 compare the predicted and actual admit *rates* for each group. Admit rates are calculated by dividing the number of admits by the number of applicants, for each racial/ethnic group.

Why was this methodology chosen?

We chose this methodology because it has many advantages over other types of analysis: (1) It makes possible a single, coherent analysis of the entire applicant pool; (2) it takes a large number of factors into account simultaneously; and (3) it does so in a way that approximates their importance in a campus' admissions process.

To understand these advantages, consider, for example, a simple comparison of admit rates across racial/ethnic groups. Such a comparison would show that African American and Chicano/Latino applicants have much lower admit rates than White and Asian American applicants. While such a comparison would reveal inequities across racial/ethnic groups in students' preparation for, and access to, higher education, it would not be useful for determining whether applicants from different groups but with similar qualifications were treated equally in the admissions process.

Comparing applicants with similar qualifications adds a level of complexity to the analysis. One way to do so would be to divide the applicant pool into clusters of students with similar SAT I scores and then to compare, within each cluster, the admission rates of applicants from different racial/ethnic groups. Such a calculation would be misleading, however, because it would account for only one of the many criteria (i.e., SAT I scores) that campuses consider when selecting students. A more sophisticated analysis might compare admit rates for clusters of applicants who have similar SAT I scores *and* similar high school GPAs, but such an analysis would still omit many other important admissions criteria. Adding additional criteria to this type of an analysis becomes problematic, however, because as criteria are added the number of clusters becomes very large, and the number of applicants in many clusters becomes too small for a reliable analysis.

This is where the logistic regression and tree models come in. Logistic regression is a statistical procedure that can be used to estimate applicants' probabilities of admission. These probabilities can then be used to cluster together applicants who, despite having different characteristics, nonetheless have similar chances of admission. For example, an applicant with an SAT I score of 1400 and a high school GPA of 3.4 might have the same probability of admission – and would therefore be placed in the same cluster – as an applicant with an SAT I score of 1100 and a GPA of 3.9.

Tree models use a different strategy, relying on applicants' characteristics (rather than estimated probabilities) to divide the applicant pool into clusters of similar students. Consider, for example, a UC campus at which 64% of the applicant pool was admitted. The tree model procedure might divide this applicant pool into two groups, one with a high school GPA of 3.7 or higher (87% admitted) and the other with a GPA of less than 3.7 (41% admitted). These subgroups would then be divided again, using whatever characteristic yields the largest distinction in admit rates. Importantly, this further division might be based on a different characteristic for each subgroup: the "3.7 and above" subgroup might be further divided by a high school GPA of 4.0 and the "below 3.7" subgroup might be further divided by an SAT II third subject score of 600. This process continues until further subdivision produces no additional distinction in the admit rates.

The final set of subgroups, or “leaves” of the tree, constitute the clusters of similar students. (An example of the tree model procedure is displayed on page 6.)

Both procedures – logistic regression and tree models – can simultaneously account for large numbers of characteristics, such as test scores, grades, proposed academic discipline, socioeconomic circumstance, and quality of the high school. Furthermore, the estimates derived under each procedure account for applicants’ characteristics in the way that best explains actual admissions decisions at the campus under consideration; as a result, the procedures approximate the importance of applicants’ characteristics in these decisions. In the analyses conducted for this study, the both procedures typically predicted about 90% of admissions decisions correctly.

Why did we use two statistical models?

As stated above, the two models differ in their approach to determining which applicants are similar. We used logistic regression last year, for a similar analysis on the 2003 cohort of UC applicants, because it is more commonly used in the social (and other) sciences. The independent statistical consultant who endorsed that analysis, however, also suggested that the tree model procedure might better predict admissions outcomes, because it can more easily account for potential interactions among various characteristics in the decisions of application reviewers. (For example, a reviewer might weigh a test score more heavily for an applicant with relatively few “a-g” courses than for an applicant with many a-g courses. A tree model is more likely to reflect such considerations than is a logistic regression.) We used both procedures this year in order to determine if they produced similar results. Both procedures are standard statistical techniques, and our implementations of both techniques have been approved by outside experts.

What are the limitations of this methodology?

Admissions processes are complex, and even the most sophisticated statistical methods can only approximate them. The logistic regression and tree model techniques estimate the impact of various criteria in the admissions process, but they cannot capture every nuance of an application reader’s deliberations. In addition, the analyses do not account for all of the criteria that these readers consider. Some of these missing criteria are quantitative in nature but were not available in UC Office of the President databases. Others are qualitative – such as academic accomplishments outside the classroom and leadership qualities – and are difficult or impossible to account for in a statistical analysis. If the distributions of these missing admissions criteria differ across racial/ethnic groups, their omission from the statistical model can cause race/ethnicity to appear to affect admissions decisions even if, in fact, it does not.

Furthermore, not only can the omission of relevant criteria from the analysis cause race/ethnicity to appear, erroneously, to affect admissions decisions, but if such an “omitted-variable bias” exists, it can affect the results for all campuses in a similar manner. To take an example, one of UC’s freshman admissions criteria is “quality of academic performance relative to the educational opportunities available in the applicant’s secondary school.” Readers evaluate applicants according to this criterion, but the data do not exist to include their assessments in the statistical model. Therefore, hypothetically, if Asian American applicants have better

opportunities, on average, than African American applicants, a statistical model that doesn't account for this may over-predict the number of Asian American students who would be admitted in a race- and ethnicity-blind process and simultaneously under-predict the number of African American admits. If such an error occurs for one campus, it is also likely to occur for other campuses that use "performance relative to opportunities" as an admissions criterion (although to varying degrees depending on the importance a campus places on this criterion), provided that a similar pattern of differences in opportunities occurs in the applicant pools for those campuses.

Even with this sophisticated methodology, therefore, it can be difficult to tell whether there are real racial/ethnic effects on admissions decisions or imperfections in the statistical models. All else being equal, small discrepancies between a group's predicted and actual number of admitted students could be due to the omission of relevant admissions criteria from the statistical model; larger discrepancies are more likely to be real effects. (UC Berkeley has recently conducted an analysis that attempts to distinguish between these two possibilities by quantifying the omitted admissions criteria and including them in a statistical model.)

Can you provide more information about the statistical models?

Yes. As mentioned above, the methodology used for these analyses simultaneously accounts for many of the criteria that each campus uses in the admissions process. These include:

High School GPA	Gender ¹
SAT I Verbal Score	High School API Decile
SAT I Math Score	Maximum Education Level of Parent (7 categories)
SAT II Writing Score	Income Level (4 categories)
SAT II Math Score (level 1 or 2)	Academic Preparation Programs (UC and non-UC)
SAT II Third Exam Score	Proposed Academic Discipline (4 categories)
ELC designation	Honors and A-G Course Counts

For some campuses, additional variables were available. For UC Irvine, the specific school to which the student applied (out of 10 possible) replaced the proposed academic discipline criterion, and an academic ranking was included. For UC Santa Barbara, a set of academic disciplines that more closely reflects that campus' admissions process (chemical engineering and computer science, electrical and mechanical engineering, and all other) replaced the proposed academic discipline criterion, and a within-high-school academic context criterion was added. For UCLA, counts of high scores on Advanced Placement or International Baccalaureate exams as well as indicators for rural school, disadvantaged neighborhood, single parent, and applicant older than 25 years, were included. For UC Berkeley, which ran its own analyses, the statistical model included many additional criteria:

- Fully honors-weighted GPA and unweighted GPA (following campus practices, the other campus' models used an honors-weighted GPA capped at eight honors courses)

¹ Gender was not used in the admissions process at any campus, but it was included in the statistical models.

- Separate within-school percentile ranks for each test score (except the third SAT II exam), both GPAs, and the numbers of honors and a-g courses
- Whether or not the applicant attended a public high school
- Whether or not the applicant came from a single-parent family
- A finer categorization of academic preparation programs (4 categories)
- The college to which the student applied (4 categories), rather than academic discipline
- Whether or not the applicant was a California resident
- Whether or not the applicant was admitted via a process, specific to the Berkeley campus and known as “augmented review,” in which the campus solicits additional information for about ten percent of its applicants

In addition to these differences in the criteria used to model admissions at each campus, there were also some differences in the populations of students considered. For most campuses, non-California residents were excluded from the analysis because they are evaluated according to different standards. UC Berkeley was the exception to this, however, since a relatively large proportion of its applicants come from outside California. (Berkeley accounted for the inclusion of non-resident applicants by including an indicator for California residency in its statistical model.) Similarly, since the vast majority of ineligible applicants are automatically denied admission, and since those who are admitted typically have a unique circumstance that would not be captured by the statistical model, applicants who were not UC-eligible were excluded from the models.^{2,3} (UC Berkeley was again the exception, although this made no difference to the results of the analysis.) At both UC Berkeley and UCLA, applicants who may have received special consideration for being athletes were excluded from the analysis, again because these applicants are evaluated according to different standards, and they constitute a significant share of the applicant pools at these campuses. For all campuses, applicants who withdrew their applications or had their applications cancelled were excluded from the analysis.

###

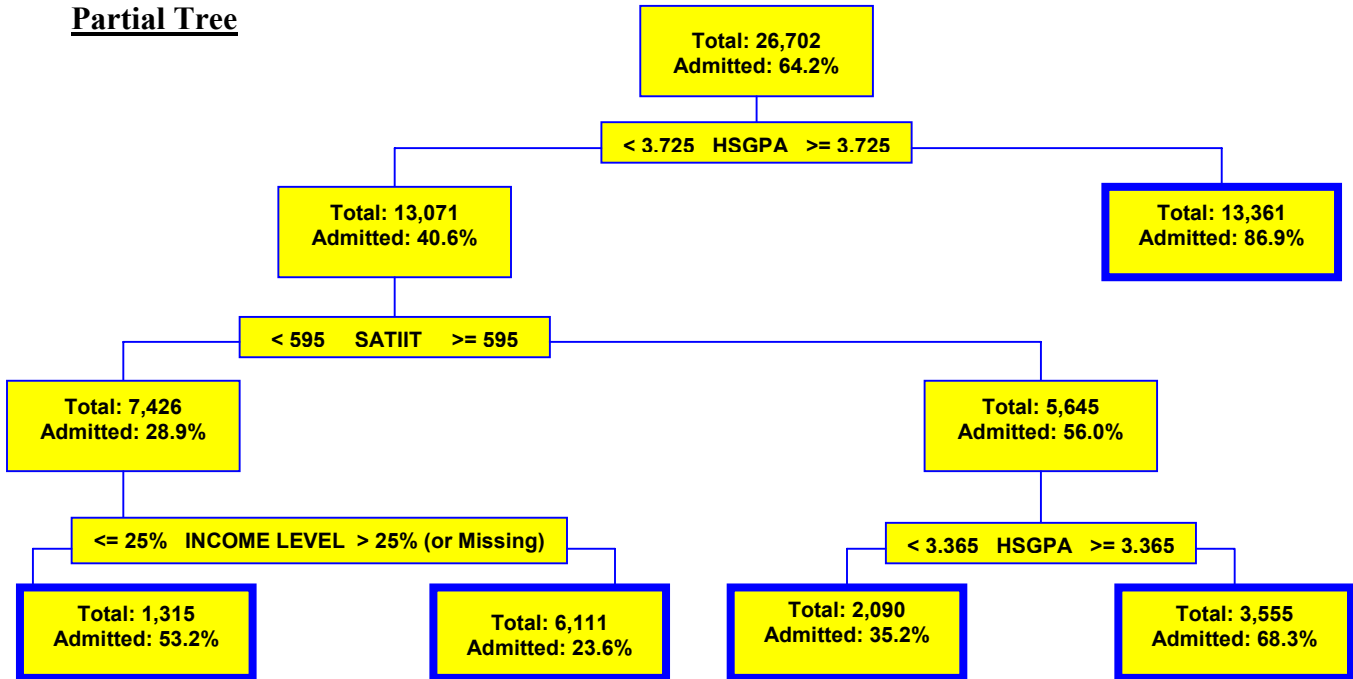
(See the following page for an example of a “tree model”.)

² Since students apply for UC before completing their senior year, campuses cannot perfectly distinguish eligible from ineligible applicants. For purposes of this analysis, and since all eligible applicants are offered admission to at least one UC campus, applicants were considered eligible if they were regularly admitted to any UC campus. This is why, even though UC Riverside admitted all eligible applicants in 2003, their actual admit rate from among eligible applicants is slightly less than 100%: Riverside may have deemed an applicant to be ineligible while another campus judged them eligible and admitted them.

³ Note that, since ineligible applicants were excluded from the analysis, admit rates presented here will differ from those reported elsewhere.

Tree Model Example (See explanation on page 2.)

Partial Tree



Full Tree

