

**UNITED STATES DISTRICT COURT FOR
THE DISTRICT OF MASSACHUSETTS
BOSTON DIVISION**

STUDENTS FOR FAIR ADMISSIONS, INC.,

Plaintiff,

v.

PRESIDENT AND FELLOWS OF HARVARD
COLLEGE (HARVARD CORPORATION),

Defendant.

Civil Action No. 1:14-cv-14176-ADB

**Declaration of
Justin McCrary, Ph.D.**

On Behalf of Defendants

July 30, 2015

I, Justin McCrary, hereby state under penalty of perjury:

I. INTRODUCTION AND QUALIFICATIONS

1. I am an economist with expertise in statistical methods, economic modeling, labor economics, law and economics, and antitrust. I received my A.B. in Public Policy from Princeton University in 1996. After working at National Economics Research Associates in White Plains, New York, and the Federal Reserve Bank of New York from 1996-1998, I began my Ph.D. in Economics at the University of California, Berkeley (“Berkeley”), completing the degree in June 2003 with field specializations in labor economics and econometrics. I then spent close to five years as Assistant Professor in the Gerald R. Ford School of Public Policy and the Department of Economics at the University of Michigan. While at Michigan, I taught introductory statistics and advanced microeconomic theory to M.P.P. students, and advanced econometric theory to Ph.D. students. I became an Assistant Professor of Law at Berkeley in January 2008 and was promoted to Professor in July 2010. While at Berkeley, I have taught courses on introductory, intermediate, and advanced statistics to J.D. students, L.L.M. students, and Ph.D. students; on law and economics to J.D. students as well as undergraduates; on business law to J.D., L.L.M., and M.B.A. students; and on labor economics to Ph.D. students.

2. In addition to my post as Professor, I am the Founding Director of D-Lab, the Social Sciences Data Laboratory at Berkeley. At D-Lab, I lecture and advise graduate students and faculty regarding high-performance computing, statistical software, statistical and econometric techniques, and research design.

3. From September 2009 until July 2014, when I began to direct the D-Lab, I co-directed the Law and Economics Program at Berkeley Law with Bob Cooter and Dan Rubinfeld (2008-2011) and with Bob Cooter and Eric Talley (2012-2014).

4. Since 2008 I have co-directed the Economics of Crime Working Group of the National Bureau of Economic Research (“NBER”). The NBER is the preeminent professional association of economists in the world, with approximately 1,300 members worldwide. I was invited to become a Faculty Research Fellow of the NBER in 2006 and remained in that position until 2012, when I was invited to become a Faculty Research Associate.

5. My research spans a diverse range of topics, including econometric and statistical methodology, education, employment discrimination, antitrust, crime, fertility, financial markets, income inequality, and monetary policy. Many of my articles have been published in leading economics, econometrics, and statistics journals, such as the *Review of Economics and Statistics* and the *Journal of Econometrics*. In addition, I have written or co-written three papers that were published in the top economics journal in the world, the *American Economic Review*, and have co-edited a book, *Controlling Crime: Strategies and Tradeoffs*, published by the University of Chicago Press. Over the years, my research has been supported by the University of Michigan; the University of California, Berkeley; the MacArthur Foundation; the NBER; the National Institutes of Health; the National Science Foundation; and the Robert Wood Johnson Foundation.

6. I am frequently asked to review articles for leading economics, econometrics, and statistics journals, including *Econometrica*, the *American Economic Review*, the *Quarterly Journal of Economics*, the *Journal of Political Economy*, the *Review of Economic Studies*, the *Review of Economics and Statistics*, and the *American Law and Economics Review*. Since coming to Berkeley Law, I have also been asked to comment on empirical papers submitted to law reviews and to peer-reviewed law journals, including the *California Law Review*, the *Law and Society Review*, the *Journal of Law and Economics*, and the *Journal of Empirical Legal Studies*.

7. I am currently a signatory to an Intergovernmental Personnel Agreement between the Equal Employment Opportunity Commission (“EEOC”) and the University of California, Berkeley. The EEOC has asked me to analyze its data regarding the racial and gender composition of the workforce of public and private employers. I receive no monetary compensation for this work.

8. My consulting experience has spanned a wide range of industries and markets. For example, I have previously analyzed the extent to which alleged collusive behavior among health care providers affected prices; the extent of infringing sales in a patent lawsuit pertaining to pharmaceuticals; the potential anti-competitive implications of a proposed telecommunications merger; damages associated with an alleged price-fixing conspiracy in the corrugated packaging industry; damages associated with an alleged price-fixing conspiracy in several prominent high-technology product markets; and damages associated with an alleged price-fixing conspiracy in the sale of retail gasoline. In addition, I am currently a consultant for the California Attorney General, tasked with analyzing the data systems maintained by the Attorney General’s office. I have also been asked to assess the extent to which those data point to differences in criminal justice outcomes between different racial groups.

9. Finally, I am frequently asked to speak on the use of statistical methodologies in empirical legal studies and for the past four summers have given day-long lectures for the week-long Causal Inference Workshop and its more advanced version, the Advanced Causal Inference Workshop, both organized by Bernie Black (a Professor of Law and Business at Northwestern University) and Matthew McCubbins (a Professor of Law and Political Science at Duke University).

10. A copy of my curriculum vitae, including a list of previous testimony and depositions, is included as Appendix A.

11. I am being compensated at my standard billing rate of \$750 per hour. I have been assisted in this matter by staff of Cornerstone Research, who worked under my direction. In addition to my direct compensation for this work, I receive from Cornerstone Research a portion of the amount that it bills for work supporting me. Neither my compensation from Defendant nor my compensation from Cornerstone Research is in any way contingent or based on the content of my opinion or the outcome of this or any other matter.

12. The statements made herein are based on my personal knowledge and upon information made available to me by Defendant's counsel and by staff of Cornerstone Research who were working under my direction.

II. BACKGROUND AND ASSIGNMENT

13. On July 16, 2015, Plaintiff filed a motion to compel the production of a preliminary sample of 6,400 application files ("Plaintiff's Motion"). I understand that the motion was filed partly in response to Defendant's proposal to produce a sample of 160 application files in conjunction with an electronic database containing information about freshman undergraduate applicants to Harvard. In support of Plaintiff's Motion, Professor Peter Arcidiacono submitted an expert declaration ("Arcidiacono Decl."). In his declaration, Professor Arcidiacono opined that a sample of 6,400 application files would be necessary in order to evaluate whether Harvard's admissions process discriminates against Asian-American applicants.¹ Professor

¹ Arcidiacono Decl. ¶¶ 27-36.

Arcidiacono further opined that Harvard's proposed sample of 160 application files would be insufficient.²

14. I have been asked to review the Arcidiacono Declaration, Plaintiff's Motion, and Harvard's electronic Admissions Office database (the "Database") in order to evaluate whether the Database (in conjunction with a sample of 160 application files) would be sufficient for the statistical analysis and modeling described by Professor Arcidiacono.

15. I submit this Declaration in support of Defendant's Opposition to Plaintiff's Motion.

III. SUMMARY OF OPINIONS

16. I have experience estimating statistical models of discrimination and understand the extant academic literature on discrimination. I also understand the types of data necessary for estimating statistical models.

17. My understanding is that Harvard is proposing to produce detailed information from the Database, for one or more years, for applicants for freshman admission. That information is comprehensive and detailed and is sufficient for the statistical analyses and modeling described by Professor Arcidiacono.

18. The Database contains data for the full universe of freshman applicants—that is, approximately 37,000 applicants each year—with several hundred fields of information in the system. For the year of data that I reviewed, for example, I understand that there are more than 900 fields in the Database. In particular, the Database contains the kinds of information identified by Professor Arcidiacono as important to the analyses he describes. Moreover, the information in the Database will be sufficient for that analysis even if Harvard redacts from the Database information that would directly identify the applicant; the applicant's family members;

² Arcidiacono Decl. ¶¶ 37-41.

and other third parties such as individual Harvard alumni interviewers, high school teachers and counselors, and others who recommended the applicant for admission to Harvard.

19. Professor Arcidiacono has not provided a reason, and I am not aware of any reason, why a sample of 6,400 application files, which is a mere subset of the universe or population of all applicants in the Database, would allow for him to perform the statistical analysis he has described more reliably than would the full Database. In fact, for the statistical analyses identified by Professor Arcidiacono, the Plaintiff's sampling method would be less reliable than analyzing the universe of applicant information contained in the Database. Furthermore, the universe or population of all applicants in the Database can be produced at a substantially lower cost than producing a sample of 6,400 application files.

20. Moreover, Harvard's proposal to also produce 160 applicant files (80 selected by SFFA and 80 selected by Harvard) in addition to the Database is more than sufficient for Professor Arcidiacono to assess whether the Database contains all the necessary information relevant for his statistical analyses.

21. I understand that the Plaintiff has argued that the proposed sample of 6,400 applications is reasonable because, at about 4 percent of the total population of applications, it is smaller in size than typical samples produced in comparable cases. From a statistical point of view, however, it is irrelevant whether or not the requested sample is modest in size relative to samples produced in other litigation. The Database includes the entire population of freshman applicants and can be produced at a substantially lower cost than a sample of 6,400 application files. Thus, the Database is preferable to any sample – regardless of size – for the analysis proposed by Professor Arcidiacono.

IV. ANALYZING A POPULATION IS PREFERABLE TO ANALYZING A SAMPLE OF THE POPULATION

22. Given Professor Arcidiacono's focus on the importance of statistical sampling, it is helpful to start with a brief overview of the purpose of statistical sampling. Any time a researcher has easy and inexpensive access to data on the *whole* population, there is no need to conduct statistical sampling. Typically, as Professor Arcidiacono and I agree, sampling is performed when collecting data is costly. In such a context, collecting a sample is less costly than collecting data on the whole population for the simple reason that a sample has fewer observations than the population. Statistical sampling selects a subset of individuals from a given population (here, the relevant population is students who applied for freshman admission to Harvard College) in such a way that the characteristics of the sample are similar to those of the population.³ A statistical sample can be used in conjunction with statistical assumptions to draw inferences regarding the population from which the sample is drawn. In summary, statistical sampling is commonly used because the relevant information for the entire population is not available or it is too costly to collect information about the entire population.

23. This fact is recognized by introductory statistics textbooks and treatises on the use of statistics in legal settings. Professor Arcidiacono and I agree on this point. As he states: "In statistical analysis, sampling relates to the selection of a subset of individuals from [a] statistical population to estimate characteristics of the whole population. Analyzing the whole population is preferable but tends to be costly."⁴

³ There are different types of statistical samples, including random samples and stratified random samples, among other types.

⁴ Arcidiacono Decl. ¶¶ 18-19

24. A simple example helps illustrate the point. Consider a population of 100 individuals, half of whom are men and half of whom are women. A researcher who did not know the fraction of the population that is female might seek to estimate this fraction using a sample of, say, 30 individuals. Though unlikely, it is possible that a random sample of 30 individuals from this population would contain 20 women and 10 men. Thus for such a random sample, the researcher's best guess about the female proportion of the population would be 66.7%.

25. Estimates based on random samples are uncertain, in the sense that different random samples would yield somewhat different estimates of the same underlying quantity: another random sample of 30 individuals from our population of 100 might contain 16 women (53.3%) instead of 20 (66.7%). For this reason, estimates derived from analysis of a sample are inherently uncertain. Statisticians and economists quantify the degree of uncertainty by reporting what is known as a confidence interval. A confidence interval is a range, such as 0.5 to 0.8, such that there is a quantifiable degree of confidence that the true fraction of the population who are women is in that range. Such confidence intervals are justified under particular statistical assumptions.

26. If one were to analyze the entire population, there would be no uncertainty of the type described above. The researcher would know that 50% of individuals in the population are female and there would be no associated confidence interval surrounding the estimate. If analyzing the entire population is no more difficult, intrusive, or costly than analyzing a sample, it is always best to analyze the population, because doing so eliminates the uncertainty in estimating the population statistic of interest.

27. Thus, to the extent that the analysis contemplated by Professor Arcidiacono can incorporate information from the full Database of applicants, there would be absolutely no reason to prefer using a sample of application files, no matter how large.

V. HARVARD'S DATABASE IS SUFFICIENT FOR THE ANALYSIS PROFESSOR ARCIDIACONO PROPOSES

28. As noted above, I have experience estimating statistical models of discrimination and am knowledgeable about the academic literature on discrimination. In my Ph.D. dissertation, I studied discrimination empirically in the context of litigation against police departments. This work was later published in the *American Economic Review* in 2007.⁵ I also draw on this expertise in connection with my work for the EEOC and the California Attorney General.

29. In teaching statistics courses at the University of Michigan and the University of California, Berkeley, and labor economics courses at the University of California, Berkeley, I routinely draw on examples of statistical models used to assess the extent to which there is evidence that might be consistent with discrimination. I also draw extensively on the literature on discrimination in teaching law and economics at the University of California, Berkeley, and have supervised Ph.D. student dissertations focusing on discrimination. As an expert in the use of statistical and econometric methods, I understand the types of data necessary for estimating statistical models such as those proposed by Professor Arcidiacono.

30. Professor Arcidiacono proposes using a sample of 6,400 application files to examine “whether application files are systematically scored differently on the basis of race.”⁶ He further elaborates on the analysis that he intends to undertake: “With the raw files in hand, we can code

⁵ McCrary, J., “The Effect of Court-Ordered Police Hiring Quotas on the Composition and Quality of Police,” *American Economic Review*, Volume 97, Number 1, March 2007.

⁶ Arcidiacono Decl. ¶ 28.

the various factors Harvard describes as important in determining application subscores (for example, creating an indicator variable for whether the student was a valedictorian). We can then use regression analysis to see, for example, whether Asians received lower subscores conditional on the factors that Harvard describes as important for that subscore.”⁷ As I discuss below, the Database is sufficient for this analysis.

31. Professor Arcidiacono also discusses using the sample of application files to score more subjective factors such as “the level and extent of” participation in extracurricular activities, so that a model predicting admissions decisions can properly account for such factors.⁸ In making these arguments, Professor Arcidiacono overlooks a much simpler, and from a statistical perspective more rigorous, solution: the use of Harvard’s Database.

32. While I understand that Professor Arcidiacono has not yet had the opportunity to review the variables included in the Database, I have looked at these variables carefully. The Database that Harvard proposes to produce contains rich, detailed information on *every* applicant for freshman admission to Harvard. In fact, the Database includes several hundred data fields for each applicant.

33. Importantly, the Database provides information on the two factors that Professor Arcidiacono has specifically identified as important.

34. First, the Database provides information relevant to determining whether an applicant was the “*valedictorian*”⁹ of his or her class. In particular, it includes information regarding the

⁷ Arcidiacono Decl. ¶ 28.

⁸ Arcidiacono Decl. ¶ 34.

⁹ I understand that “*valedictorian*” may be defined differently for different candidates, because the honor can be defined differently by different secondary schools. Some candidates may have received the honor at the time of applying, and others may expect to receive the honor after applying. My understanding is that if the applicant disclosed among his or her honors that he or she is the *valedictorian*, that designation is captured in the

class rank of some applicants (as I understand it, those for whom their guidance counselor filled out this information in the application materials), and it includes the applicant's self-reported list of high-school honors and achievements.

35. Second, the Database contains extremely detailed information about each applicant's extracurricular activities. For example, for each extracurricular activity that an applicant reports on his or her application (up to a maximum of twelve), the Database contains fields identifying the type of activity (for example, School Newspaper/Journalism), the applicant's role (for example, Editor-in-Chief), the years during which the applicant was involved in the activity, whether the applicant's involvement was year-round or only during the school year, and the number of hours per week and number of weeks per year that the applicant devoted to the activity.

36. These are just two examples of the detailed information available in the Database. There are additional types of information in the Database that could be relevant for Professor Arcidiacono's analyses. For example, the Database includes:

- Numerous measures of academic success while in high school (*e.g.*, GPA, class rank, scores on the ACT, SAT I, SAT II, and TOEFL, AP courses taken and scores on AP tests, honors reported by the applicant (such as National Merit Scholar), and the level of the honor (such as national vs. school));
- Measures of intended areas of focus while at Harvard (*e.g.*, intended major, intended career, and intended graduate school plans);
- Extensive information on each applicant's family (*e.g.*, parents' marital status and education level, and siblings' age and education level, and whether the applicant's parent(s) and/or siblings attended Harvard College);
- Information about the applicant's and the applicant's family's financial situation, (*e.g.*, parents' occupation and employer, whether the applicant applied for financial aid, and whether they paid their application fee);

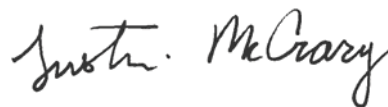
database. Because many students do not know if they are going to be valedictorian at the time of the application, this information is often not disclosed in the application and therefore not captured in the database.

- Cultural and demographic information (*e.g.*, race and ethnicity, languages spoken by the applicant, and proficiency level for each language); and
- Ratings assigned by Harvard admissions officers and alumni interviewers (which includes admissions officer ratings for academics, extracurricular activities, athletics, and personal qualities).

37. Given this extensive record for each applicant, it is my opinion that the Database is sufficient for the analyses described in the Arcidiacono Declaration and preferable to the sampling method proposed by the Plaintiff and Professor Arcidiacono. I am not aware of any reason why a sample of 6,400 application files would allow for the analysis proposed by Professor Arcidiacono to be conducted more reliably than would Harvard's Database. In fact, for the analyses that Professor Arcidiacono has described, his sampling method would be less reliable than analyzing the full universe of applicant information contained in the Database.

38. To the extent the Plaintiff would like to ascertain that the Database contains all the necessary information for their statistical analysis, Harvard's proposal to also produce 160 applicant files (80 selected by SFFA and 80 selected by Harvard) is more than sufficient to make this assessment.

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct. Executed on July 30, 2015.



Professor Justin McCrary,
Ph.D.