Dissertation Review: Experiments on Labor Market Discrimination by Daviti Jibuti

This dissertation provides evidence on labor market discrimination by conducting three field experiment. The first experiment uses a correspondence experimental design to examine whether there is discrimination against applicants to jobs in the financial sector with visible tattoos and whether the discrimination varies by observed characteristics of the applicant, firm or job. Using a sample of 782 observations and estimation using linear probability models, the study shows that applicants without visible tattoos have a 13 percentage point higher callback rate. This result is remarkably robust to inclusion of a variety of controls in the linear probability models. The second chapter describes a similar correspondence experiment that examines labor market discrimination against applicants with visible tattoos in the IT sector. The sample size is 800 observations. Estimates from linear probability models show 7-9 percentage point lower callback rates for applicants with visible tattoos.

Both studies attempt to distinguish between statistical and taste-based discrimination. In the second experiment, the fictitious applicants also vary on skill levels signalled quite directly. Nevertheless, the evidence in favor of statistical discrimination is not statistically significant so the study concludes that tattooed applicants may face taste-based discrimination.

The third chapter describes a study of hiring bias in a context in which participants are required to make hiring decisions for a mathematically intensive task. The total number of participants is 2216. Using a discrete choice econometric model, the study finds hiring biases against Black and less attractive workers, and for Asian and female workers. The experimental design also includes variation in the number of candidates to choose from. This feature allows for a statistical test of the impact of the size of the choice set on hiring decisions. The study finds hat increasing the size of the choice set reduces the likelihood that a Black or Latino candidate is chosen and increases the likelihood that an Asian candidate is chosen.

I found that dissertation addresses a novel and, in big picture, important topic. The experimental design in each of the studies is sound and the statistical analysis is competently done. The results in each chapter are substantive and the dissertation, as a whole, advances

our knowledge substantively. It is also is well written. Therefore, I believe that the dissertation satisfies all requirements for a PhD thesis in economics and is ready to be defended after the first set of editorial suggestions, below, are incorporated into the dissertation. Below those suggestions, I have some thoughts on how this work might be strengthened further and/or extended into future work.

1. I can imagine policy makers arguing that discrimination against individuals with chosen characteristics, such as tattoos, does not reach the level of importance that discrimation against non-choice characteristics like race and gender has. I think there are a number of ways to argue against this conjecture. I think the introduction should conclude with a paragraph or two that couches the policy relevance of this work (specifically the first two studies) in a broader context, i.e., by making it about more than "just tattoos".

2. p 16. "If the company rejected the candidate, the observation was considered a negative response. Finally, if the firm did not respond at all, it was considered a rejection, ..." Although this seems reasonable (and commonplace in the literature), I'm not sure the implicit assumption in making this analytic choice is completely harmless. A sentence or two justifying this approach - even if it's just citing other work - should be included.

3. I'm puzzled about the definitions of variables such as urban and West Germany in the regression specifications that give rise to results reported in Chapter 1, Table 4 and Chapter 2, Table 3. To my understanding, those variables should be collinear with region fixed effects. What am I missing? Please address so readers like me are not left puzzled about the specification.

4. Perhaps because I am not an expert in the field of correspondence studies, I have struggled to understand the difference between a correspondence study and an audit study. To help readers like me, a clarifying sentence or two in the introduction would be very helpful.

5. p 8 and elsewhere. Is the p - value "equal to" 0.0001 or is it "less than or equal to"?

6. p 12. "Another drawback of the method lies in the technique itself: as the method is limited to the job offers, I can respond to with written applications, I might exclude

some specific jobs that do not require written applications (especially low skilled jobs)." seems garbled. Please rewrite.

7. p 13. There is a hanging "Concluded".

8. p 20. The equation should include notation for covariates because most of the regression results reported have covariates in the reqressions.

9. p 32. Here, and elsewhere, there are tables that show "marginal effects at means of the Probit model". Are these really marginal effects at covariate means? Or are they means of marginal effects? If it's the former, is there a reason for that choice? Current practice is to report means of marginal effects so if effects at means are reported that choice should have some explanation.

10. The discrete choice model used in Chapter 3 is clearly a conditional logit model yet I didn't see that described anywhere in the body of the paper (only once in the appendix). The conditional logit has a long history and considerable study of its characteristics so the fact that the choice model is a conditional logit should be made clear in the body of the chapter.

11. p 73 table 2. What is "prop"? I suspect it is an abbreviation for proportion. If so, I don't understand the variable in the context of the conditional logit model. Shouldn't those variables all be indicators for whether the alternative has those characteristics?

12. It was not until a footnote in Chapter 2 that I learned about the results of the power analysis of the experimental design. This is important and should be reported more prominently in Chapters 1 and 2.

These are some thoughts for improvement of these studies post-defense and/or future work.

1. The estimated effect sizes appear to be sitting at the edge of the power curve (800 observations needed to detect an effect size of 10 percentage points). Therefore, it is not surprising that most of the interaction effects are not statistically significant,

and that some of the specifications in Chapter 2 show insignificant effects. I think it would be very interesting, and possibly illuminating, to estimate effects using a pooled sample of observations from the financial and IT sector designs. That would double the sample size and increase power. I recognize that there are differences between the two designs but there are also many similarities so assumptions required for pooling should be relatively mild.

2. The interpretation of results from the conditional logit can be enhanced a lot. It's quite possible that interesting results are hidden because the interpretation in Chapter 3 is entirely from the linear index (with minimal transformation to get odds ratios). It is possible to calculate marginal effects from such models, and compare all kinds of alternatives with each other. From Stata 16 onward, it has become much easier to do – see `asclogit`.