Identity and Bias: Insights from Driving Tests*

Revital Bar and Asaf Zussman
Economics Department, The Hebrew University of Jerusalem

March 8, 2017

Abstract

How does one's identity affect the evaluation of others? To shed light on this question, we analyze the universe of driving tests conducted in Israel during 2006-2015, leveraging the effectively random assignment of students and testers to tests. We find strong and robust evidence of both ethnic (Arab/Jewish) *in-group* bias and gender *out-group* bias: a student is 15 (11) percent more (less) likely to pass a test when assigned a tester from the same ethnicity (gender). We show that these patterns are consistent with a utility-based interpretation, along the lines of Becker's (1957) taste-based discrimination model.

JEL classification codes: J15.
Keywords: Identity, Bias, Discrimination.

*We thank Simon Jager, David Neumark, Gautam Rao and audiences at Bar Ilan University, Ben Gurion University, Hebrew University, I-Core, and the Weizmann Institute of Science for useful comments and to Hanania Afangar, Ella Dorfman, Galia Hardon, Effi Rozen, Dalit Tamari and Elena Zlocisty from the Israeli Ministry of Transport and Road Safety for their help with the data. Financial support for the project was generously provided by the I-Core Program of the Planning and Budgeting Committee at the Israel Science Foundation (grant no. 1821/12).

1 Introduction

How does one's identity affect the evaluation of others? In this paper we shed light on this question using data on driving tests. A driving test is a standard procedure designed to test a person's ability to drive a motor vehicle under normal operating conditions. Such tests are conducted in most countries around the world and serve as a requirement for obtaining a driver's license. Testers are typically government employees who are expected to assess students' driving abilities in an impartial manner. At the same time, however, testers enjoy a great deal of discretion in making their decisions, which opens the door for bias and discrimination.

Specifically, the paper studies ethnic (Arab/Jewish) and gender bias using data on the universe of driving tests conducted in Israel between 2006 and 2015. Most of our analysis focuses on private vehicle tests – in total, more than 2.5 million such tests were conducted during this period. Identification of causal effects relies on the essentially random assignment of testers and students to tests.

We find evidence of both ethnic *in-group* bias and gender *out-group* bias: a student is 15 percent more likely to pass a test when assigned a tester from the same ethnic group and 11 percent less likely to pass a test when assigned a tester from the same gender. We show that these results (1) are not driven by potential confounds such as the influence of tester characteristics (other than ethnicity and gender); and (2) are robust to various changes in the estimated equations.

We argue that the observed patterns are inconsistent with classical models of statistical discrimination (Arrow, 1972 and Phelps, 1972), which in our context would imply that when evaluating the driving abilities of individual students, testers might be influenced by rational and accurate perceptions regarding the distribution of driving skills of students from different ethnicities and genders. First, since tests are thirty-minutes long, testers have enough time to directly observe students' driving abilities and do not need to rely on perceptions of cross-group differences in these abilities. Second, testers conduct

thousands of tests each year and thus should have relatively uniform perceptions regarding the driving skills of students from different groups. Third, the statistical discrimination model would predict that more experienced testers would be better able to estimate individual students' driving abilities and therefore would need to rely less on statistical inference. We find no evidence of such a relationship in our data.

Another potential interpretation of the results is that they are driven by implicit bias, i.e. a bias that operates at a level below conscious awareness and characterizes split-second decisions (Greenwald and Banaji, 1995). This type of bias has received a lot of scholarly attention in recent years.[1] While it is possible that implicit bias plays a role in determining the tester's "first impression" of the student, we believe that since testers evaluate students for half an hour, the effect of implicit bias on test outcome should be small.

The leading alternative to statistical discrimination and implicit bias is the taste-based discrimination model (Becker, 1957). The key element in this model is that agents incur different levels of utility from contact with members of different groups.[2] We argue that our results are consistent with such a utility-based model. In driving tests, testers sit next to students and interact with them. This interaction might affect test outcomes by influencing tester's utility, with a higher utility level leading to a higher pass rate. In simple terms, our results seem to imply that testers reward members of groups they enjoy interacting with. We provide three tests to support this interpretation.

First, we explore whether the extent of bias in driving tests is correlated with measures of prejudicial attitudes. The analysis focuses on ethnic bias, capitalizing on the fact that inter-ethnic relations in Israel exhibit considerable spatial and temporal variation. Similar to Charles and Guryan (2008), prejudicial attitudes are measured using the extent of public support for laws

[1] Banaji and Greenwald (2013) provide a thorough discussion and review of the literature on implicit bias. Examples of such research include Bertrand, Chugh and Mullainathan (2005) and Milkman, Akinola and Chugh (2015).

[2] Akerlof and Kranton (2000), who study the role of identity in economic life, provide an alternative utility-based model of discrimination. In their model, for example, the "distaste" of men for working with women can be understood as emanating from a loss in male identity when women work in a man's job.

banning inter-racial marriages. Consistent with the taste-based interpretation, we find positive and strong spatial and temporal associations between bias and prejudicial views.

Second, we argue that if bias is indeed driven by the different levels of utility testers derive from interaction with members of different groups during the test, it is natural to assume that this effect would decline with physical distance between testers and students. To explore this hypothesis, we replicate our analysis of bias using data on driving tests for motorcycle licenses, where the student and the tester drive different vehicles and are thus not in close proximity.[3] We find no evidence of bias in motorcycle tests.

Third, building on previous literature showing that short term variation in weather affects decision-making through cognitive biases, we examine the association between daily temperatures and bias in driving tests. We find evidence that higher temperatures are associated with greater ethnic (but not gender) bias.

The literature on discrimination and bias is extensive. Most of it focuses on the identity of the subject of evaluation, e.g. studying discrimination against job applicants from specific ethnic groups. Our paper is most closely related to a strand in the literature which examines the effect of a *match* between the identity of the evaluator and the identity of the subject of evaluation. Researchers use this approach for two main purposes. First, in some situations there are no objective measures of performance, ability, qualification etc., which makes it impossible to argue that differences in outcomes between members of different groups are due to discrimination. In these situations, examining the effect on outcomes of a match in identity between the evaluator and the subject of evaluation allows researchers to credibly establish the existence of discrimination when assignment is random. Second, this approach allows the researcher to better understand the mechanisms underlying observed bias and in particular to disentangle taste-based from statistical discrimination. The idea is that if bias is statistical in nature, its extent should

[3]Since there is only one female tester conducting motorcycle tests, in this case too we focus on ethnic bias.

not vary with the evaluator's identity.

An important dichotomy within this literature is between studies that rely on (lab or field) experiments and those that rely on naturally occurring data. While experiments give researchers greater control, in many cases they suffer from well-known weaknesses such as the fact that decision makers are not professional, groups are artificially-generated and stakes are low. The use of naturally occurring data overcomes these difficulties.

Recent examples of research that examines the effect of a match between evaluator's and subject's identities and relies on naturally occurring data include papers exploring bias in: judicial decision making (Shayo and Zussman (2011 and forthcoming), Anwar, Bayer, and Hjalmarsson (2012) and Depew, Eren and Mocan (2016)); refereeing in academic journals (Abrevaya and Hamermesh (2012)) and in sports (Price and Wolfers (2010) and Parsons et al. (2011)); policing (Anwar and Fang (2006), Antonovics and Knight (2009) and West (2016)) and lending decisions (Fisman, Paravisini and Vig (2017)).

Our paper makes three contributions to the existing literature. First, the fact that we are able to study two biases simultaneously allows us to show that the type of identity examined matters for the direction of bias. While the result of ethnic in-group bias is in line with the typical finding in the literature, to our knowledge, this paper is the first to show evidence of gender out-group bias.

Second, the results suggest that utility considerations may play a more prominent role in generating bias than is usually assumed. While the context of driving tests may seem somewhat esoteric, this insight may be relevant in other, more economically important, contexts. In the situation we study, testers serve as screeners and are not expected to interact with the students after the test. This resembles a hiring situation in which the job interview is conducted by a human resources officer. Although in many cases the officer will not have any contact with the candidate in the future, the utility he or she derives from the interaction during the interview may affect the hiring decision. In case the interview is conducted directly by the employer – which in essence is the situation portrayed in Becker's employer discrimination model – the

effect might be even stronger. This is because the employer may derive utility both from the interaction during the interview but also from the prospects for future interactions.[4] A similar mechanism may be at play in other contexts when decision making is accompanied by face to face interaction, for example when loan officers screen applicants.

The third contribution is methodological. Most of the recent empirical literature examines discrimination using a correspondence study methodology. Our finding regarding the effect of physical proximity on bias implies that correspondence studies – which, by design, involve no physical interaction between the relevant parties – may (1) underestimate the *extent* of bias and (2) fail to correctly identify the *source* of bias (underestimating the role of taste-based discrimination while giving too much weight to statistical discrimination).[5]

Two limitations of the difference-in-differences approach we employ in this study are worth noting. First, we are able to estimate only relative rather than absolute levels of bias against certain groups. Suppose, for example, that in addition to the utility-based considerations we have emphasized so far, both male and female testers incorrectly believe that female students are less able drivers than male students. In this case, we would only be able to pick up the effect of tastes but not the effect of stereotypes. Second, we are unable to say which group of testers is biased and what is the direction of bias. In the case of ethnicity, for example, we cannot determine whether Jewish testers, Arab testers or both are biased. Moreover, it is impossible to determine whether testers from a specific group are biased in favor of students from their own group or biased against students from the other group.

The rest of the paper is structured as follows. Section 2 provides details on the institutional context. Section 3 describes the datasets we use in the analysis and provides summary statistics. In Section 4 we show results of balancing

[4] Our argument is consistent with studies that find that the race of the hiring agent or manager is associated with the racial composition of new hires in firms. See, for example, Stoll, Raphael and Holzer (2004) and Giuliano, Levine, and Leonard (2009).

[5] Correspondence studies largely replaced audit studies as the experimental method of choice in uncovering discrimination. While audit studies do involve face-to-face interaction, they suffer from well-known weaknesses (see, for example, Neumark (2016)).

tests, outline the empirical strategy and provide the main results concerning ethnic and gender bias. Section 5 addresses potential confounds and shows results of robustness checks. In Section 6 we explore possible interpretations of the results. Section 7 concludes.

2 Driving Tests in Israel

In this section we describe the institutional context in which driving tests are conducted, focusing on private vehicle tests.

2.1 Geographical Structure

The Israel Ministry of Transport and Road Safety (MOT) divides the country into 4 regions: (1) Tel Aviv and Center; (2) Haifa and North; (3) Be'er Sheba and the Negev; and (4) Jerusalem and South. Each of these regions contains several testing centers; overall, there are 43 centers. Each MOT tester and each driving school – and through it each driving teacher and student – is associated with one of these regions.[6]

2.2 Students

The first step in the journey to obtain a driving license starts when the student arrives at an MOT-certified facility and is issued an official form (called the "green form"). The form, which is specific to the type of driving license the student wishes to obtain (e.g. private vehicle or motorcycle), initially includes the student's photograph and personal details. Students must later have the form signed by an optometrist and a family doctor certifying that they are physically fit to drive. Students then have to pass a driving theory test and take lessons in an MOT-certified driving school.

[6] It is important to emphasize that MOT driving tests are taken by citizens and permanent residents of Israel. This includes Israelis residing in Jewish West Bank settlements and Arab residents of East Jerusalem but excludes Palestinians residing in the West Bank.

Students can first take the theory test when they turn 16 and 3 months. The theory test can be taken in six different languages, including Hebrew and Arabic. The 40 minute long test is comprised of 30 multiple choice questions. Students must answer at least 26 correctly in order to pass the test. They may retake the test as many times as they need to.

When students are 16 and 6 months old, they can start taking driving lessons. Students must take at least 28 driving lessons – each lasting 40 minutes – before they can take the MOT driving test. This requirement may be reduced by the teacher to 20 lessons under special circumstances, e.g. in case the student already holds a driving license for a different type of vehicle. Our conversations with MOT officials indicate, however, that most students take more than the required minimum number of lessons.[7]

When the teacher believes that the student is prepared to take the MOT driving test, she first assigns him to an "internal test". Internal tests are conducted by the professional manager of the driving school (driving schools usually have several teachers but may also have only one, in which case the teacher is also the manager of the school). If the student fails the internal test, he needs to take additional driving lessons. Once the student passes the internal test, he is eligible to take the MOT driving test (the minimum age for taking the MOT driving test is 16 and 9 months). The student must be tested in the same region to which his driving school belongs.

2.3 Teachers

In order to become an MOT-certified driving teacher, one must be at least 21 years old, have completed 12 years of education, hold a driving license for at least 3 years and have no criminal record. As a first step in the selection and training process of driving teachers, eligible candidates undergo rigorous assessment by an external human resources firm. Only about 20 percent of candidates obtain a passing score in this assessment. These candidates have to then take a practical driving test, where they are expected to exhibit out-

[7] The price of one 40-minute long driving lesson varies between NIS 100 and NIS 150 (approximately $US 30-40).

standing driving skills. The next step is attending a 680 hours driving teacher course (this takes approximately 2 years). The vast majority of those who start the course complete it successfully and receive a driving teacher certificate from the MOT. This certificate is relevant for teaching only for a private vehicle license. Teachers who wish to teach driving for other types of licenses, need to undergo additional training.

2.4 Testers

The minimum requirements for becoming an MOT driving tester are similar to those for becoming a teacher, except that testers must be at least 25 years old. Candidates undergo assessment by the same human resources firm as teachers and, like them, also need to pass a driving test. The professional course for testers is somewhat longer than that of teachers. Certified driving teachers who want to become testers need to take a shorter version of the tester course. The MOT uses a competitive tender process to recruit the most suitable candidates out of those who have successfully completed the course. The recruitment process is region-specific. Selected candidates undergo additional training, where they join experienced testers in conducting actual tests. Once this additional training period is over, the candidates are tested by the head tester in their region. Upon passing this last hurdle, they receive their tester certificate and can start testing.

2.5 Assignment

The assignment of testers to tests is based on computerized, region-specific, waiting lists. Students enter these lists once they pass the theory test. Those who pass the driving test drop out of the list, while those who fail remain in it.

Before the beginning of each month, the MOT compares – for each region separately – the number of students waiting to be tested to the number of available tests (the latter figure is based on the availability of testers in that month). This yields region-specific ratios which are then used to allocate a

specific number of tests to each teacher. Thus, for example, if the region-specific ratio is 4, a teacher in this region with 20 students in the waiting list will be allocated 5 slots. A test slot is defined by a test center, date and time. Crucially, the MOT does not inform the teachers about the identity of the tester in each slot.

The four MOT region offices construct a weekly work plan for each tester, detailing in which test centers they will work each day. For example, in a certain week, a specific tester from the Be'er Sheba and the Negev region might be assigned to work in Be'er Sheba on Sunday and Tuesday, in Netivot on Monday and in Sderot on Wednesday through Friday. These assignments are revealed to the testers a week in advance. Only when the tester shows up for work in the morning, is he provided with a work schedule for that day specifying the name of the driving school for each time slot. Under no circumstances are testers allowed to deviate from this schedule.[8] With this work schedule in hand, the tester approaches a designated parking area and locates the vehicle of the specific driving school assigned to him (the name of the school appears on the car). The identity of the student is revealed to the tester (and vice versa) *only* when the tester enters the car.

The main objective of the MOT assignment procedure is to make sure that testers will not be able to choose whom to test and students (and their teachers) will not be able to choose whom to be tested by. This implies that the assignment of students to testers is effectively random. In other words, within a test area, the likelihood of being assigned a tester from a given ethnicity or gender is the same for all students. We later use balancing tests to check whether assignment is indeed random.

2.6 Tests

A test begins when the tester enters the car. On the dashboard are waiting for him the student's identification card and green form as well as a receipt

[8] The only exception occurs when a student is assigned a tester who has already failed him at least 3 times in the past. In this (extremely rare) case, the student can ask to be assigned to a different tester.

for payment for the test.[9] The tester fills the student's details in his daily schedule form, wishes her good luck and instructs her to start driving. Tests are allocated between 25 and 30 minutes.

At the end of the test, after leaving the car, the tester fills out a detailed test evaluation form. The form is divided into three main sections, each containing more than a dozen criteria: (1) control of the vehicle (e.g. control of the steering wheel); (2) traffic (e.g. merging into traffic); and (3) the road (e.g. turning right or left). The tester marks only those criteria where the student demonstrated poor performance. Based on these marks, the tester decides whether the student passed or failed, writes a short explanation for the decision in the evaluation form and records the decision in the green form. The tester then returns the evaluation form and the green form to the MOT test center office. The forms are later distributed back to the teachers and, through them, to the students.

How do testers decide whether to pass or fail a student? Although testers are well trained and have detailed testing guidelines, assessing the driving skills of students based on dozens of criteria is very much subjective. Moreover, there is no official formula for aggregating the separate marks into a single outcome. Taken together, these facts imply that testers have a lot of discretion in making the pass/fail decision.[10] In fact, in our data the average pass rate per tester – for testers who conducted at least 1,000 tests – varies greatly: it is 26 percent at the 5th percentile and 62 percent at the 95th percentile.[11]

[9] Payment for the test has two components. The first is a fee paid to the MOT while the second compensates the driving teacher for the use of his vehicle in the test. During the period examined here, the total payment amounted to about $US 100.

[10] We further note that students' ability to successfully appeal testers' decisions is very limited. Based on our conversations with MOT officials, only 2-3 percent of failures are appealed, and out of these, 90 percent are rejected after a conversation between the tester who conducted the test and the regional head tester. In the remaining cases, students are allowed to retake the test with the head tester (with no additional costs to them).

[11] The large variability in pass rates across testers was noted in an October 2016 report by the State Comptroller of Israel on the operation of the MOT's Licensing Division. The report recommended that measures would be taken to reduce testers' discretion and increase uniformity in pass rates.

3 Data

The MOT provided us with 3 datasets. The first contains information on the universe of driving tests conducted between June 2006 and September 2015. Each observation includes the following fields: test outcome (pass/fail), scrambled student identification number, scrambled tester identification number, test date, test area, number of theory tests, the current driving test number and the type of driving license the test is for. The dataset contains information on over 3 million tests, of which 81 percent are for private vehicle licenses and 8 percent are for motorcycle licenses. The rest are tests for licenses for buses, trucks, tractors etc.. Our analysis focuses on private vehicle tests.[12]

The second dataset contains information on the students who took these tests. Each observation contains the following fields: scrambled identification number, first name, gender, birth year, locality of residence, zip code within this locality, type of license for which the student was tested and identification keys for driving school and teacher. The dataset contains information on more than a million students.

The third dataset has information on the driving testers who performed the tests in the first dataset. Each observation has the following fields: scrambled identification number, first name, gender, birth year, locality of residence and zip code within this locality. The dataset covers 236 testers for private vehicle licenses.

To deduce the ethnicity of students and testers we use the following two-step procedure. The first step uses first names to assign ethnicity, building on the fact that Arabs and Jews in Israel have very different naming conventions. This approach has been used in previous research dealing with ethnicity in Israel, e.g. Shayo and Zussman (2011) and Zussman (2013 and 2016). Specifically, we utilize a dataset derived from the Israeli population registry which provides, separately for each gender, the probability that a given first name belongs to an Arab citizen. We identify a name as Arab if it is at least twice

[12] To explore the sources of bias, in Section 6 we additionally utilize the data on motorcycle tests.

as popular among Arabs than it is among Jews, and as Jewish if it is at least twice as popular among Jews than it is among Arabs. This first step enables us to assign ethnicity to 91 percent of students and 93 percent of testers.

To assign ethnicity to the remaining students and testers, in the second step we rely on the fact that there is a very high degree of residential ethnic segregation in Israel. The population of most localities is either all-Arab or all-Jewish and the population of integrated localities, such as Jerusalem and Tel Aviv, is ethnically segregated by neighborhood. To code ethnicity based on place of residence, we use three datasets from the Israeli Central Bureau of Statistics. The first classifies localities as either Arab, Jewish or integrated. The second provides, for each statistical area (sub-neighborhood), the ethnicity to which the majority of residents belong. The third maps zip codes into statistical areas.[13] Thus, we first classify students and testers as Arab if they reside in Arab localities, and as Jews if they reside in Jewish localities. This assigns ethnicity to 90 percent of those whose ethnicity we were not able to ascertain using first names. We use the data on the main ethnicity in each statistical area to assign ethnicity to the remaining students and testers (who live in ethnically-integrated localities). Overall, our two-step procedure enables us to assign ethnicity to all testers and to 99 percent of students; the remaining students are excluded from the analysis.[14]

Finally, we merge the 3 files to create one dataset. The students dataset was merged with the tests dataset using a combination of scrambled student identification number and license type. The testers dataset was merged with the tests dataset using the scrambled tester identification number. To illustrate the structure of the merged dataset, consider the example of a student who

[13]There are more zip codes than statistical areas. In most cases, a zip code is entirely contained in a single statistical area. In some cases, however, a zip code is divided by two statistical areas. In those cases, we follow a "majority rule": we assign the zip code to the statistical area that has most addresses.

[14]The main reason we assign ethnicity first using names and only then by relying on locality and zip code, is that we have the names of all students and testers while information on residence is missing for a non-trivial share of students' and testers' observations. In sub-section 5.2 we show that our results are robust to reversing the order of the ethnicity identification procedure.

passed his third test for a private vehicle license. This student would have 3 observations in this dataset: 2 for the failed tests and one for the successful test.

3.1 Summary statistics

Panel A of Table 1 shows the distribution of private vehicle tests across MOT regions by the ethnicity of students and testers. We note several interesting patterns in the data. Seven percent of tests were conducted by Arab testers while 29 percent of tests were taken by Arab students. The share of cross-ethnicity tests (where the tester and the student belong to different ethnic groups) is 30 percent. This share exhibits significant variation across MOT regions: it is 18 percent in the Tel Aviv and Center region and 50 percent in the Haifa and North region. This variation stems from the fact that the Arab population of Israel is not uniformly distributed across the different regions of the country.

[Table 1]

Panel B of Table 1 shows the distribution of tests across MOT regions by the gender of students and testers. Eight percent of tests were conducted by female testers while 55 percent of tests were taken by female students. The share of cross-gender tests is 55 percent; as expected, this share does not vary much across regions.

Table 2, Panel A, provides summary statistics on students. Column 1 shows means (and standard deviations) for all students while columns 2-3 and 5-6 provide means (and standard deviations) for different ethnic groups and genders. About 25 percent of students are Arab and roughly 50 percent are female (column 1). Students are young: the average age is about 23 (the median, not reported in the table, is 19). The average number of driving tests is 1.9 for Arab students and 1.6 for Jewish students; the corresponding figures are 1.8 for female students and 1.6 for male students. Arab students take on average 3.1 theory tests while Jewish students take only 1.9. Both male and female students take about 2.2 theory tests on average.

[Table 2]

Summary statistics for testers are provided in Panel B of Table 2. About 9 percent of testers are Arab and roughly the same share of testers is female. The average age of testers is 54, with Arab testers being 5 years younger than their Jewish colleagues; female testers are on average about 6 years younger than male testers. To capture the possibility that workload might influence testers' decisions, in the analysis below we control for the number of tests each tester conducted on the day of the test. Testers in the different groups conduct on average between 9 and 12 tests per day.

4 Ethnic and Gender Bias

In this section we explore whether a student is more (or less) likely to pass a test when assigned a tester from his or hers own ethnic group or gender. Our ability to credibly identify such biases crucially depends on the assumption that the assignment of students to testers is random.

Table 3 shows the results of balancing tests examining this issue. We first analyze balance with respect to tester ethnicity. For each student characteristic, column 1 reports the mean and standard deviation of this characteristic for students assigned to Arab testers, column 2 shows the corresponding statistics of this characteristic for students assigned to Jewish testers and column 3 tests whether the means are equal.

[Table 3]

Results in the first row indicate that the share of students who are Arab is 45.8 percent when the tester is Arab and only 28.2 percent when the tester is Jewish, yielding a large and statistically significant difference in means of 17.6 percentage points. This difference is not surprising given the fact that, as mentioned in conjunction with Table 1, Arabs tend to live in specific areas of the country. Indeed, when we test for the equality of means while controlling for test area fixed-effects (column 4, first row), the difference declines to

0.1 percentage points and becomes statistically insignificant. The next rows replicate this analysis for student gender, age, and the number of driving and theory tests. While the differences in means for some of these characteristics are statistically significant, their magnitudes are miniscule.[15]

In columns 5-8 of Table 3 we conduct balancing tests with respect to tester gender. In this case, the raw means of all characteristics of students assigned to male and female testers are quite similar (columns 5-7). After adjusting for test area fixed-effects (column 8), the differences in means, while statistically significant, are again extremely small.

Taken as a whole, the results of the balancing tests show that the assignment of students to testers seems to be effectively random with respect to tester ethnicity and gender.

4.1 Ethnic Bias

Figure 1 displays student pass rates by tester and student ethnicity. When the tester is Jewish (left two columns), the pass rate is 42.5 percent for Jewish students but only 32.7 percent for Arab students. In itself, this 9.8 percentage points difference does not indicate the existence of ethnic bias. It is possible, for example, that on average, Arab students arrive to the test less prepared than Jewish students. If this was the only difference between Arab and Jewish students, we would expect a similar cross-ethnicity difference in pass rates when the tester is Arab. In fact, however, we observe that when the tester is Arab (right two columns), the pass rate is 33.6 percent for Jewish students

[15] To gain perspective, the results of these balancing tests can be compared to those performed by Shayo and Zussman (2011), who explore whether the assignment of cases to judges in Israeli small claims courts is balanced with respect to judge ethnicity. While none of the differences in observable case characteristics they test for turns out to be statistically significant, the magnitude of some of the differences in means is non-negligible. For example, after adjusting for court fixed-effects, the difference between the share of Arabs among plaintiffs assigned to Arab judges and the share of Arabs among plaintiffs assigned to Jewish judges is 1.3 percentage points. This difference is an order of magnitude larger than the one we report above for the assignment of Arab students to Arab and Jewish testers. A major difference between the current paper and Shayo and Zussman (2011), which leads us to reject the null hypothesis of equality of means for some of the characteristics, is that the number of observations is more than 1,500 times larger in the current study.

and 33.0 percent for Arab students (a 0.6 percentage points difference). The difference in these differences, of 9.2 percentage points, is the raw estimate of the extent of in-group bias (Appendix Table A1 reports this difference-in-differences analysis in more detail). It is crucial to note that in the absence of an objective measure of driving ability, it is impossible to determine whether Jewish testers are biased toward Jewish students, Arab testers are biased toward Arab students, or some combination of the two.[16]

[Figure 1]

Next, we explore ethnic bias econometrically. We start by estimating the following basic specification:

$$\begin{aligned} Pass_{ijat} &= \alpha_0 + \alpha_1 ArabStudent_i + \alpha_2 ArabTester_j \\ &\quad + \alpha_3 ArabStudent_i * ArabTester_j + \delta_a + \epsilon_{ijat} \end{aligned} \tag{1}$$

where $Pass_{ijat}$ is an indicator for passing the test for student i, tested by tester j, in test area a, on date t; *ArabStudent, ArabTester* and the interaction term *ArabStudent*∗*ArabTester* are indicator variables; δ_a is a test area fixed-effect; and ϵ_{ijat} is an error term clustered within tester. This specification allows for differences in pass rates across ethnic groups that are not necessarily due to bias. Specifically, the equation captures possible differences in driving abilities between Arab and Jewish students (α_1) and possible differences in leniency between Arab and Jewish testers (α_2). Our interest is in the coefficient α_3, which captures the extent of bias.

Column 1 of Table 4 presents the results from estimating equation (1). Controlling for test area fixed-effects, we find that when the tester is Jewish, Arab students are 5.5 percentage points less likely to pass the test than their Jewish peers. For Jewish students, the likelihood of passing the test is 4.8 percentage points lower when the tester is Arab. As stressed above, these results

[16] Of course, it is also possible that both groups of testers are biased against Arab students (or in favor of Jewish students), but Arab testers are less biased than their Jewish colleagues.

in themselves do not indicate the existence of in-group bias. The coefficient for the interaction variable, which captures in-group bias, is estimated at 7.1 percentage points and is highly statistically significant. Considering that the overall pass rate is 39.3 percent, the bias seems quite large: a student is 18 percent more likely to pass a test when assigned a tester from his or hers own ethnic group.

[Table 4]

We next gradually augment equation (1) with additional controls. The most elaborate specification is the following:

$$\begin{aligned} Pass_{ijat} &= \alpha_0 + \alpha_1 ArabStudent_i + \alpha_3 ArabStudent_i * ArabTester_j \\ &\quad + \delta_a + \theta_t + \mu_1 S_{it} + \mu_2 T_{jt} + \gamma_j + \epsilon_{ijat} \end{aligned} \quad (2)$$

where θ_t is a set of controls for year, month and day of week of the test; S_{it} is a set of student characteristics – female indicator, age in test, driving test number (i.e. number of previous driving tests + 1) and number of theory tests; T_{jt} is a set of time varying tester characteristics – age in test and number of tests conducted by the tester on the same day; and γ_j is a tester fixed-effect.[17]

The inclusion of these additional controls lowers the estimate of in-group bias from 7.1 percentage points in column 1 to 5.9 percentage points in column 5. The latter estimate is still large (about 15 percent of the mean pass rate) and highly statistically significant.

4.2 Gender Bias

Figure 2 displays student pass rates by tester and student gender. When the tester is male, the pass rate is 44.1 percent for male students but only 35.7 percent for female students, an 8.3 percentage points difference. When the tester is female, the pass rate is 44.7 percent for male students but only

[17] Note that adding tester fixed-effects to the estimated equation makes the inclusion of tester characteristics that are not time varying, i.e. ethnicity and gender, redundant.

31.9 percent for female students, a 12.9 percentage points difference. This indicates the existence of gender *out-group* bias of a substantial magnitude: 4.5 percentage points or 11 percent (Appendix Table A2 reports this difference-in-differences analysis in more detail). As in the case of ethnic bias, there is no way to determine whether male testers favor female students, female testers favor male students or some combination of the two.[18]

[Figure 2]

In Table 5 we explore gender bias econometrically, relying on the approach used in equations (1) and (2) but replacing the ethnicity variables with the corresponding gender variables. Controlling for test area fixed-effects, we find that when the tester is male, female students are 7.5 percentage points less likely to pass the test than male students. For male students, the likelihood of passing the test are similar regardless of the gender of the tester. The out-group bias estimated with the basic model (column 1, third row) is 4.4 percentage points. This estimate drops only slightly to 4.2 percentage points with the full set of controls (column 5).

[Table 5]

We next examine ethnic bias and gender bias simultaneously, using the following basic specification:

$$\begin{aligned} Pass_{ijat} &= \alpha_0 + \alpha_1 ArabStudent_i + \alpha_2 ArabTester_j \\ &\quad + \alpha_3 ArabStudent_i * ArabTester_j \\ &\quad + \beta_1 FemaleStudent_i + \beta_2 FemaleTester_j \\ &\quad + \beta_3 FemaleStudent_i * FemaleTester_j \\ &\quad + \delta_a + \epsilon_{ijat} \end{aligned} \tag{3}$$

[18] The fact that female students have a much lower pass rate than male students, regardless of tester gender, may raise the possibility that the performance of females in the test is affected to some degree by a "stereotype threat" (Steele and Aronson, 1995). This refers to a situation in which people feel themselves to be at risk of conforming to stereotypes about their social group. This mechanism may be especially relevant in the current context since the perception that women do not drive as well as men is quite prevalent in Israel.

The coefficients capturing ethnic bias (α_3) and gender bias (β_3) presented in column 1 of Table 6 are almost identical to those presented in column 1 of Tables 4 and 5. The estimated ethnic *in-group bias* is 6.9 percentage points and the estimated gender *out-group bias* is 4.5 percentage point. We next augment this basic specification with the regular set of controls. The most elaborate specification is the following:

$$\begin{aligned} Pass_{ijat} &= \alpha_0 + \alpha_1 ArabStudent_i + \alpha_3 ArabStudent_i * ArabTester_j \\ &+\beta_1 FemaleStudent_i + \beta_3 FemaleStudent_i * FemaleTester_j \\ &+\delta_a + \theta_t + \mu_1 S_{it} + \mu_2 T_{jt} + \gamma_j + \epsilon_{ijat} \end{aligned} \quad (4)$$

where all the variables are as defined in equation (2). Using the most elaborate specification, ethnic bias is estimated at 5.9 percentage points (15 percent) and gender bias is estimated at 4.2 percentage points (11 percent). Both estimates are highly statistically significant. Overall, the results presented in Table 6 imply that the two biases are to a large degree orthogonal to each other.[19]

[Table 6]

5 Confounds and robustness

5.1 Potential confounds

In this subsection, we address two potential confounds. First, as documented in panel B of Table 2, Arab testers differ from their Jewish colleagues in their characteristics (for example, Arab testers are on average 5 years younger).

[19] To further explore this issue, we conduct the following exercise. First, we estimate equation (4) – with the necessary modifications – for each tester separately, focusing on testers who conducted at least 1,000 tests during the entire period (leaving us with 176 of 236 testers). Second, we regress the coefficient for *ArabStudent* (α_1) on the coefficient for *FemaleStudent* (β_1), controlling for tester ethnicity and gender. Consistent with the argument that the two biases are orthogonal, we find that the correlation between α_1 and β_1 is close to zero in value and statistically insignificant.

This may confound interpretation of the results if, for example, regardless of tester ethnicity, older testers treat Arab students differently than their younger colleagues. We address this concern by adding to equation (2) interactions between the *ArabStudent* indicator and tester characteristics other than ethnicity. Results are in Appendix Table A3.

To facilitate comparison, in column 1 we replicate the results from column 5 of Table 4. Columns 2 to 4 show that the additional interaction terms are for the most part statistically insignificant and, more importantly, that the estimate of ethnic in-group bias maintains its size and statistical significance. This pattern remains when including in the regression all the interactions simultaneously (column 5).

We perform an analogous exercise to rule out the possibility that our estimate of gender out-group bias is driven by differences in mean characteristics between male and female testers (for example, female testers are on average 6 years younger than male testers). Results, presented in Appendix Table A4, show that the interactions between student gender and tester characteristics other than age are statistically insignificant. The estimate of gender out-group bias varies between 3.1 and 4.2 percentage points and remains statistically significant throughout.

So far we have interpreted the observed differences in outcomes across groups as reflecting tester behavior. A potential confounding factor – which is shared by most studies in the relevant literature – is the possibility that student behavior during the test is endogenous to the ethnicity or gender of the tester. For example, students may perform poorly in the test when assigned a tester from the opposite ethnic group or from the same gender.[20]

To address this concern, we rely on the following insight. While students may react to the ethnicity or gender of the tester, they are not likely to react to tester characteristics that are not observed by them. At the same time, some of these characteristics may influence tester behavior. A notable example for

[20] In a recent paper, Glover, Pallais and Pariente (2016) provide evidence of such endogenous reaction. They examine the performance of cashiers in a French grocery store chain and find that manager bias negatively affects minority job performance.

such a characteristic is whether the tester resides in an integrated locality. It is very unlikely that a student would be able to infer during the test in which type of locality the tester resides, but there is reason to believe that residence in integrated localities may be correlated with views concerning Arab-Jewish relations that in turn may influence test outcomes. Specifically, according to the well-known "contact hypothesis" (Allport, 1954), cross-group contact – that in the current context is inherent to residence in integrated localities – would work to reduce prejudice.

To explore this issue, we compare outcomes in tests conducted by testers from integrated versus non-integrated localities. Because in our data only two Arab testers reside in integrated localities and one of the two conducted only 11 tests, we limit the analysis to Jewish testers. We start by estimating the following basic model:

$$\begin{aligned} Pass_{ijat} &= \alpha_0 + \alpha_1 ArabStudent_i + \alpha_2 TesterInt_j \\ &\quad + \alpha_3 ArabStudent_i * TesterInt_j + \delta_a + \epsilon_{ijat} \end{aligned} \tag{5}$$

where *TesterInt* is an indicator for (Jewish) testers residing in integrated localities. The other variables are defined as before. In the next step we gradually augment this specification with the regular set of controls. Our interest is in the coefficient α_3, which captures the difference in outcomes for Arab students when they are tested by testers residing in integrated rather than by testers residing in non-integrated (Jewish) localities.

Results of the analysis suggest that, consistent with our original interpretation, test outcomes are significantly influenced by tester behavior (Table 7). Moreover, the results are also consistent with the predictions of the "contact hypothesis": we find that Arab students are 2-3 percentage points more likely to pass the test when tested by Jewish testers residing in integrated rather than all-Jewish localities.[21]

[21] Admittedly, the additional analysis does not completely rule out a possible role for endogenous student behavior. For example, it is possible that Jewish testers from integrated localities behave in a way that makes Arab students feel more comfortable during the test.

[**Table 7**]

An additional confound relates to the fact that driving tests are conducted in Hebrew, possibly generating difficulties in communication between testers and students who do not share the same native language. We believe that this is unlikely to be a major issue. Given that all testers pass a rigorous training and selection process in Hebrew, there is strong reason to believe that Arab testers are perfectly fluent in Hebrew. While it is likely that some Arab students are not as fluent in Hebrew as the testers, it is hard to believe that this would create a serious barrier given the simplicity of the driving instructions provided by the testers.

5.2 Robustness

We next provide several tests for the robustness of our results. One concern might be that the results are driven by a single tester or a single test area. To address this concern, we repeatedly estimate equation (4), each time dropping one tester or one test area. Our estimates of ethnic bias and gender bias barely change.[22] Appendix Figures 1A and 1B further illustrate that there are no individual testers whose biases are particularly notable. Figure 1A displays the coefficient for *ArabStudent* obtained when regressing, for each tester separately, test outcome on an *ArabStudent* indicator and the regular set of controls. Testers are ordered from left to right based on the size of the coefficient. The figure illustrates that the value of the coefficient varies smoothly across testers, with Arab testers concentrated on the right side. Figure 1B similarly shows that the coefficient for *FemaleStudent* varies smoothly across testers, with Female testers concentrated on the left side.

Recall that to identify the ethnicity of both students and testers, we first rely on names and then on location information. We identify a name as Arab if it is at least twice as popular among Arabs than it is among Jews, and as

We view this as another form of tester bias.

[22] The estimate of ethnic bias varies between 0.052 and 0.070, and the estimate of gender bias varies between -0.048 and -0.040. In all cases the estimates remain highly statistically significant.

Jewish if it is at least twice as popular among Jews than it is among Arabs. We conduct two robustness checks of this procedure. In the first, we replicate the analysis of ethnic bias (column 5 of Table 4) using a stricter criterion: we identify a name as Arab (Jewish) if it is at least three times as popular among Arabs (Jews) than it is among Jews (Arabs). In the second check, we identify ethnicity first using location information and then by relying on names. Results are robust to both changes (Appendix Table A5).

Students' performance in the test may reflect differences in teaching styles and other characteristics of driving teachers. To control for these differences, we augment equation (4) with a driving teacher fixed-effect (the driving teacher identifier is available for the vast majority of students in our data). The results, presented in Appendix Table A6, are robust to this change.

The performance of students in the test may obviously also depend on a host of unobserved student characteristics (e.g. coordination skills). To control for such factors, we leverage the fact that many students need to take more than one test to obtain their driving license and estimate the following equation:

$$\begin{aligned} Pass_{ijat} &= \alpha_0 + \alpha_3 ArabStudent_i * ArabTester_j \\ &\quad + \beta_3 FemaleStudent_i * FemaleTester_j \\ &\quad + \delta_a + \theta_t + \mu_1 S_{it} + \mu_2 T_{jt} + \gamma_j + \lambda_i + \epsilon_{ijat} \end{aligned} \tag{6}$$

where λ_i is a student fixed-effect and all the other variables are as defined above.[23] In this analysis, identification of ethnic bias comes from students who were tested by testers from different ethnic groups (these student took 482,012 tests) and identification of gender bias comes from students who were tested by testers from different genders (these students took 575,402 tests). Results are presented in Table 8. Following the addition of student fixed-effects, the estimate of ethnic in-group bias drops by about a third, while the

[23]Note that adding student fixed-effects implies dropping from the estimated equation student characteristics that do not vary over time (i.e. ethnicity and gender).

estimate of gender out-group bias slightly increases; both estimates remain highly statistically significant.

[Table 8]

6 Interpretation

In this section we examine possible sources for the observed biases. Like most of the literature in economics, we focus on the distinction between the two leading models of discrimination: statistical and taste-based.

6.1 Statistical discrimination

Statistical discrimination means that when assessing attributes of specific agents from different groups, decision makers take into account cross-group differences in the distributions of those attributes. The canonical example of statistical discrimination describes a hiring situation in which an employer uses information about differences in the average productivity levels of different racial groups when evaluating individual job candidates from these groups. In the current context, statistical discrimination would imply that when evaluating the driving abilities of individual students, testers might be influenced by perceptions regarding the driving skills of, for example, Arab versus Jewish students.

We argue that it is unlikely that our results concerning ethnic in-group bias and gender out-group bias are driven solely or even mainly by statistical discrimination. First, testers do not need to rely at all on group averages since they can directly observe, for 30 minutes, the individual student's driving abilities – the only relevant attribute for making a well-informed pass/fail decision. In this regard, it is important to note that testers are not expected to forecast whether, upon receiving the license, the student will be a safe driver. Rather, according to a top MOT official we spoke with "the aim of the test is to ensure that the student can drive the vehicle from point A to point B without causing harm".

Nevertheless, one might argue that as long as the signal obtained in the test is not perfectly informative, testers would still need to rely on their ex-ante perceptions regarding the driving skills of students from different groups. Where do those perceptions come from? Our data show that testers conduct on average roughly 1,800 tests per year, with very little variation across testers' ethnicities and genders. This implies that all testers should have uniform perceptions regarding the driving skills of students from different groups and therefore should reach similar decisions. Our results stand in sharp contrast to this prediction.[24]

Additionally, the statistical discrimination model predicts that more experienced testers would be better able to estimate individual students' driving abilities and therefore would need to rely less on group averages. We would thus expect bias to decline with experience. In Table 9 we test this hypothesis, using age as a proxy for tenure. Column 1 replicates the results from estimating equation (4) for the sake of comparison. In column 2 we add interactions between tester age and the following variables: *ArabStudent, ArabTester* and the interaction term *ArabStudent∗ArabTester.* In column 3 we redo this analysis using interactions between tester age and the variables *FemaleStudent, FemaleTester* and the interaction term *FemaleStudent∗FemaleTester.* Column 4 includes both sets of interactions simultaneously. The results show that neither ethnic nor gender bias diminishes with tester experience, although given that the coefficients of interest are not tightly estimated, we cannot rule out this possibility.

[Table 9]

[24] A distinction that is sometimes made with respect to statistical discrimination is between true and false stereotypes. The patterns we observe can be rationalized only with false stereotypes that differ by tester identity, e.g. male testers view females as worse drivers (relative to males) than female testers do. Such a pattern of divergent beliefs is part of what social psychologists have long considered as a major manifestation of in-group bias and is inconsistent with statistical discrimination.

6.2 Taste-based discrimination

The leading alternative to statistical discrimination is Becker's taste-based discrimination model. The key element in this model is that some agents incur different levels of utility from contact with members of different groups. Returning to the canonical hiring situation described above, a white employer facing two equally-productive job candidates, one black and the other white, would prefer to hire the latter because he incurs a disutility from interacting with the former. In the context of driving tests, a Becker-style model would argue that testers would be more likely to pass students from a specific group because they derive higher utility from interacting with them.[25]

We argue that our results seem consistent with some form of taste-based discrimination and provide three tests to support this interpretation: the first relates variation in the extent of bias to measures of prejudice; the second examines how bias is affected by physical proximity between testers and students; and the third investigates whether bias depends on short term weather variation.

6.2.1 Prejudice and bias

Our test of the relationship between bias in driving tests and prejudice focuses on ethnic bias and capitalizes on the fact that inter-ethnic relations in Israel vary considerably over space and time. To measure prejudicial attitudes, we follow the approach taken by Charles and Guryan (2008). They use data on wages and on attitudes – taken from the General Social Survey – to provide evidence for Becker's employer discrimination model in the United States. Specifically, Charles and Guryan show that the black-white wage gap is larger in areas characterized by stronger prejudicial views (or racial animus). Their main measure of such views is the extent of public support for laws banning

[25] Obviously, the driving test context is different from the hiring situation in many ways. One difference is that a discriminatory employer sacrifices profits to indulge his tastes while a discriminatory tester does not incur a direct cost for being biased. However, testers may still bear some cost for acting in a discriminatory manner. The cost may be either psychological (e.g. because testers wish to view themselves as impartial) or material (e.g. because there is a risk that testers' decisions may be subject to official scrutiny at some point).

inter-racial marriages.

In Israel, no official survey asks questions of this sort. However, Zussman (2013) conducted large scale surveys to measure the attitudes of Jews towards Israeli Arabs. Among other things, the survey asked participants to report their degree of support for laws banning inter-ethnic marriages. The survey spanned the period from August 2009 to April 2011 and included 3,600 participants. Our measure of ethnic bias is thus the share of participants who support (strongly or otherwise) a ban on inter-ethnic marriages.

To conduct the spatial analysis, we first assign to each tester the sub-district in which he or she resides. We then run equation (2) separately for each sub-district (the analysis is limited to the seven out of fifteen sub-districts that have testers and students from both ethnic groups). In Figure 3 we plot the estimated bias in driving tests against the support for a ban on inter-ethnic marriages in each sub-district. We find that ethnic bias is positively correlated with prejudicial attitudes (the correlation coefficient is 0.46 and is statistically insignificant).

[Figure 3]

To measure the temporal variation in ethnic bias, we apply a rolling regression technique. Specifically, we estimate equation (2) using moving seven-quarter windows.[26] Figure 4 shows the estimated coefficients together with 95 percent confidence intervals. Ethnic bias varies considerably over time but is always positive and statistically significant.

[Figure 4]

For the seven quarters for which we have the survey data, Figure 5 plots ethnic bias in driving tests against the share supporting a marriage ban. Consistent with the hypothesis that bias is driven by prejudice, the association

[26] To illustrate, the regression centered on quarter t covers tests conducted from quarter $t-3$ through quarter $t+3$; the following regressions are centered around quarters $t+1$, $t+2$ etc..We note that at the beginning and at the end of the period analyzed, windows are by necessity shorter than seven quarters.

between the two variables is positive and strong (the correlation coefficient is 0.88 and is highly statistically significant).[27]

[Figure 5]

6.2.2 Physical Proximity

In the context we study, testers sit next to students and interact with them. By influencing the utility enjoyed by the tester during the test, this interaction might affect test outcomes, either consciously or unconsciously. In simple terms, we argue that it is possible that testers reward members of groups they enjoy interacting with, i.e. members from their own ethnic group and from the opposite gender.

If indeed bias is driven by the different levels of utility testers derive from interaction with members of different groups, it seems natural to assume that this effect would depend on the physical distance between testers and students. Specifically, we argue that the (relative) disutility incurred by testers from interacting with members of a "disliked" group would decline with physical distance.

To test this hypothesis, we replicate our analysis of bias using data on motorcycle tests. The institutional details concerning motorcycle tests are almost identical to those concerning private vehicle tests. Importantly, as in the case of private vehicle tests, testers are not able to choose whom to test and students are not able to choose whom to be tested by. The key difference between the two types of tests is that in motorcycle tests, the student and the tester drive different vehicles and are thus not in close proximity. Since there is only one female tester conducting motorcycle tests, we focus again on ethnic bias. Appendix Tables B1-B3 provide summary statistics and balancing checks for motorcycle tests.[28]

[27] It may seem natural to leverage spatial and temporal variation in fatalities from Palestinian terrorism to estimate the effect of inter-ethnic tensions on the extent of bias in driving tests. However, using this approach is not feasible in the current context, because the period analyzed here was characterized by few such fatalities.

[28] The results of the balancing tests indicate that the assignment of students to testers is

In Table 10 we compare the extent of bias in private vehicle tests and in motorcycle tests. Column 1 replicates the results obtained previously from estimating equation (2) for private vehicle tests (column 5 of Table 4). It is important to note that some testers conduct only private vehicle tests while others conduct both private vehicle tests and motorcycle tests (i.e. none of the testers conduct only motorcycle tests). To make sure that we compare the extent of bias across vehicle types for the same group of testers, in column 2 we restrict the analysis of ethnic bias in private vehicle tests to testers who conduct both types of tests. The estimated bias is slightly smaller than that estimated for all the testers (4.5 vs. 5.9 percentage points) but is still highly significant. Column 3 shows the results from estimating bias in motorcycle tests. We find no evidence of ethnic bias in these tests, which is consistent with the hypothesis that bias depends on physical proximity.[29]

[Table 10]

6.2.3 Temperature and Bias

To provide further support for our argument that testers' decisions are not reached solely through rational calculations but are rather driven by utility considerations and emotions, we examine the association between temperature and bias. Previous research has shown that short term variation in weather affects decision-making through cognitive biases, e.g. Saunders (1993), Conlin, O'Donoghue and Vogelsang (2007) and Simonsohn (2009). Capitalizing on this insight, we examine the effect of daily variation in temperature on tester behavior. Specifically, we augment our database with daily data on the maximum temperature measured in the weather station closest to each test center. We then examine, separately for each ethnic group and for each gender,

effectively random. As in the case of private vehicle tests, while some of the differences are statistically significant, their magnitudes are small.

[29] An alternative explanation for the finding that ethnic bias is lower in motorcycle tests could be the difference in the salience of ethnicity across the two types of tests. Salience is lower in motorcycle tests because: (1) testers and students interact face-to-face only at the beginning and at the end of the test and (2) testers and students wear helmets during the test.

whether the variation in temperatures affects test outcomes. For example, for Jewish testers we estimate the following equation:

$$\begin{aligned} Pass_{ijat} &= \alpha_0 + \alpha_1 ArabStudent_i + \alpha_2 Temperature_{at} \\ &+\alpha_3 ArabStudent_i * Temperature_{at} \\ &+\delta_a + \theta_t + \mu_1 S_{it} + \mu_2 T_{jt} + \gamma_j + \epsilon_{ijat} \end{aligned} \tag{7}$$

where all the variables (other than temperature) are as defined above. Our interest is in the coefficient α_3, which captures the effect of temperatures on test outcomes of Arab students relative to Jewish students. We follow the same procedure, with the necessary modifications, for Arab testers, male testers and female testers.

The results in Table 11 indicate that Jewish testers are affected by variation in temperatures while Arab testers are not (columns 1- 2): when the maximum daily temperature rises by ten degrees Celsius, the probability of passing a test conducted by a Jewish tester decreases by 0.26 percentage points for Jewish students but by three times as much (0.80 percentage points) for Arab students. Variation in temperatures does not seem to affect gender bias.

[Table 11]

What can account for the effect we find for Jewish testers? One possibility, consistent with some previous findings in the literature, e.g. Card and Dahl (2011), is that the rise in temperatures leads to greater aggression which, in our context, is expressed mainly towards members of the out-group.

7 Conclusion

This paper sheds light on the effect of identity on the evaluation of others by studying ethnic and gender bias in driving tests in Israel. The analysis utilizes data on the universe of tests conducted between 2006 and 2015 and exploits the effectively random assignment of testers and students to tests. We find

evidence of both ethnic *in-group* bias and gender *out-group* bias: a student is 15 percent more likely to pass a test when assigned a tester from the same ethnic group and 11 percent less likely to pass a test when assigned a tester from the same gender.

Our paper makes three contributions to the literature on discrimination and bias. First, the fact that we study two biases simultaneously allows us to show that the type of identity examined matters for the direction of bias. While the result concerning ethnic in-group bias is in line with the typical finding in the literature, to our knowledge, this paper is the first to show evidence of gender out-group bias.

Second, the results suggest an important role for utility considerations in generating bias. We argue that decisions made by professional screeners, operating under a non-discriminatory norm, may be influenced by the utility they derive from face-to-face interaction with the subjects of evaluation. This insight may be relevant in other, more economically important, contexts.

Relatedly, we note that most of the recent empirical literature examines discrimination using a correspondence study methodology. Our finding regarding the effect of physical proximity on bias implies that correspondence studies may (1) underestimate the *extent* of bias and (2) fail to correctly identify the *source* of bias – underestimating the role of taste-based discrimination while giving too much weight to statistical discrimination.

8 References

Abrevaya, Jason, and Daniel S. Hamermesh. 2012. "Charity and Favoritism in the Field: Are Female Economists Nicer (to Each Other)?" *Review of Economics and Statistics,* 94(1): 202–207.

Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics,* 115 (3):715-753.

Allport, Gordon W. 1954. *The Nature of Prejudice.* MA: Addison-Wesley.

Antonovics, Kate, and Brian G. Knight. 2009. "A New Look at Racial Profiling: Evidence from the Boston Police Department." *Review of Economics and Statistics,* 91(1): 163-177.

Anwar, Shamena, Patrick Bayer, and Randi Hjalmarsson. 2012. "The Impact of Jury Race in Criminal Trials." *Quarterly Journal of Economics*, 127(2):1017-1055.

Anwar, Shamena, and Hanming Fang. 2006. "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence." *American Economic Review*, 96(1):127-151.

Arrow, Kenneth J. 1972. "Some Mathematical Models of Race in the Labor Market." In *Racial Discrimination in Economic Life*, ed. Anthony H. Pascal, 187-204. Lexington, MA: Lexington Books.

Banaji, Mahzarin R., and Anthony G. Greenwald. 2013. *Blindspot: Hidden biases of good people.* Delacorte Press.

Becker, Gary S. 1957. *The Economics of Discrimination.* Chicago: The University of Chicago Press.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan. 2005. "Implicit Discrimination." *American Economic Review,* 95(2): 94–98.

Card, David, and Gordon B. Dahl. 2011. "Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior." *Quarterly Journal of Economics,* 126(1):103-143.

Charles, Kerwin K., and Jonathan Guryan. 2008. "Prejudice and Wages: An Empirical Assessment of Becker's *The Economics of Discrimina-*

tion." *Journal of Political Economy*, 116(5): 773-809.

Conlin, Michael, Ted O'Donoghue, and Timothy J. Vogelsang. 2007. "Projection Bias in Catalog Orders." *American Economic Review*, 97(4): 1217-1249.

Depew, Briggs, Ozkan Eren, and Naci Mocan. 2016. "Judges, Juveniles and In-group Bias." National Bureau of Economic Research, Working Paper no. 22003.

Fisman, Raymond, Daniel Paravisini and Vikrant Vig. 2017. "Cultural Proximity and Loan Outcomes." *American Economic Review,* 107(2): 457–492.

Giuliano, Laura, David I. Levine, and Jonathan Leonard. 2009. "Manager Race and the Race of New Hires." *Journal of Labor Economics,* 27(4): 589-631.

Glover, Dylan, Amanda Pallais, and William Pariente. 2016. "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores." National Bureau of Economic Research, Working Paper no. 22786.

Greenwald, Anthony G., and Mahzarin R. Banaji. 1995. "Implicit Social Cognition: Attitudes, Self-esteem, and Stereotypes." *Psychological Review*, 102(1): 4–27.

Milkman, Katherine L., Modupe Akinola, and Dolly Chugh. 2015. "What Happens Before? A Field Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway into Organizations." *Journal of Applied Psychology*, 100(6): 1678–1712.

Neumark, David. 2016. "Experimental Research on Labor Market Discrimination." National Bureau of Economic Research, Working Paper no. 22022.

Parsons, Christopher A., Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh. 2011. "Strike Three: Discrimination, Incentives, and Evaluation." *American Economic Review*, 101(4), 1410-1435.

Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review*, 62(4): 659-61.

Price, Joseph, and Justin Wolfers. 2010. "Racial Discrimination

among NBA Referees." *Quarterly Journal of Economics*, 125(4): 1859-1887.

Saunders, Edward M. 1993. "Stock Prices and Wall Street Weather." *American Economic Review*, 83(5): 1337-1345.

Shayo, Moses, and Asaf Zussman. 2011. "Judicial Ingroup Bias in the Shadow of Terrorism." *Quarterly Journal of Economics*, 126(3):1447-1484.

Shayo, Moses, and Asaf Zussman. Forthcoming. "Conflict and the Persistence of Ethnic Bias." *American Economic Journal: Applied Economics.*

Simonsohn, Uri. 2010. "Weather to Go to College." *Economic Journal*, 120(543): 270-280.

Steele, Claude M., and Joshua Aronson. 1995. "Stereotype Threat and the Intellectual Test Performance of African Americans." *Journal of Personality and Social Psychology,* 69(5): 797-811.

Stoll, Michael A., Steven Raphael, and Harry J. Holzer. 2004. "Black Job Applicants and the Hiring Officer's Race." *Industrial and Labor Relations Review*, 57(2): 267–87.

West, Jeremy. 2016. "Racial Bias in Police Investigations." Working Paper.

Zussman, Asaf. 2013. "Ethnic Discrimination: Lessons from the Israeli Online Market for Used Cars." *Economic Journal,* 123(572): F433–F468.

Zussman, Asaf. 2016. "Conflict and the Ethnic Structure of the Marketplace: Evidence from Israel." *European Economic Review*, 90: 134-145.

Table 1
Geographical Distribution of Driving Tests, by MOT Regions

Panel A: By Tester and Student Ethnicity

MOT Region	Number of Test Areas	Tester: Student:	Jewish Jewish	Jewish Arab	Arab Jewish	Arab Arab	Tests
Tel Aviv and Center	14		80.09	15.34	3.10	1.48	1,072,687
Haifa and North	14		42.02	43.47	6.75	7.75	820,404
Be'er Sheba and the Negev	10		74.15	22.54	2.58	0.74	221,382
Jerusalem and South	5		76.26	23.03	0.53	0.18	501,448
Countrywide	43		66.91	26.24	3.71	3.13	2,615,921

Panel B: By Tester and Student Gender

MOT Region	Number of Test Areas	Tester: Student:	Male Male	Male Female	Female Male	Female Female	Tests
Tel Aviv and Center	14		41.54	47.88	4.92	5.66	1,072,687
Haifa and North	14		36.68	55.06	3.31	4.96	820,404
Be'er Sheba and the Negev	10		44.42	52.66	1.30	1.62	221,382
Jerusalem and South	5		45.29	50.29	2.10	2.32	501,448
Countrywide	43		40.98	51.00	3.57	4.46	2,615,921

Notes. The table shows, for each MOT region, the share (in %) of driving tests in each combination of student and tester ethnicities (panel A) and genders (panel B).

Table 2
Summary Statistics

	Panel A: Students (N=1,097,836)						
	All students	Arab students	Jewish students	Difference	Female students	Male students	Difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Arab student	0.251	1.000	0.000	1.000	0.271	0.230	0.041***
	(0.433)	(0.000)	(0.000)	[N/A]	(0.444)	(0.421)	[0.001]
Female student	0.508	0.549	0.494	0.055***	1.000	0.000	1.000
	(0.500)	(0.498)	(0.500)	[0.001]	(0.000)	(0.000)	[N/A]
Student age in test	23.18	22.99	23.24	-0.248***	23.86	22.48	1.379***
	(9.455)	(8.02)	(9.889)	[0.019]	(9.001)	(9.851)	[0.018]
Number of driving tests	1.691	1.896	1.623	0.273***	1.801	1.578	0.224***
	(0.899)	(1.093)	(0.813)	[0.002]	(0.991)	(0.778)	[0.002]
Number of theory tests	2.194	3.147	1.875	1.270***	2.136	2.255	-0.119***
	(2.506)	(3.473)	(1.985)	[0.007]	(2.255)	(2.739)	[0.005]

Notes. Standard deviations are in parentheses in columns 1-3 and 5-6. Standard errors are in brackets in columns 4 and 7. Each entry in column 4 is derived from a separate OLS regression where the explanatory variable is an indicator for Arab student. Each entry in column 7 is derived from a separate OLS regression where the explanatory variable is an indicator for Female student. Number of driving tests is the current test number, i.e. number of previous failed tests plus one. Number of theory tests is the number of theory tests the student has taken.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Table 2
Summary Statistics

	Panel B: Testers (N=236)						
	All testers	Arab testers	Jewish testers	Difference	Female testers	Male testers	Difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Arab tester	0.085	1.000	0.000	1.000	0.095	0.084	0.012
	(0.279)	(0.000)	(0.000)	[N/A]	(0.301)	(0.278)	[0.070]
Female tester	0.089	0.100	0.088	0.012	1.000	0.000	1.000
	(0.285)	(0.308)	(0.284)	[0.070]	(0.000)	(0.000)	[N/A]
Tester age in test	53.96	49.29	54.39	-5.104***	48.26	54.52	-6.263***
	(8.324)	(6.883)	(8.326)	[1.610]	(6.716)	(8.268)	[1.544]
Number of same day tests	9.491	11.64	9.292	2.344***	11.58	9.286	2.298***
	(4.367)	(3.244)	(4.410)	[0.771]	(3.106)	(4.424)	[0.730]

Notes. Standard deviations are in parentheses in columns 1-3 and 5-6. Standard errors are in brackets in columns 4 and 7. Each entry in column 4 is derived from a separate OLS regression where the explanatory variable is an indicator for Arab tester. Each entry in column 7 is derived from a separate OLS regression where the explanatory variable is an indicator for Female tester. Number of same day tests is the total number of tests the tester conducted on the day of the observed test.

*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Table 3
Balancing Tests for the Assignment of Students to Testers, by Ethnicity and Gender

	Mean		Differences in Means Arab vs. Jewish Tester		Mean		Differences in Means Male vs. Female Tester	
	Arab tester	Jewish tester	No controls	w/ Area FE	Female tester	Male tester	No controls	w/ Area FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Arab student	0.458	0.282	0.176***	-0.001	0.302	0.293	0.008***	-0.003***
	(0.498)	(0.450)	[0.001]	[0.001]	(0.459)	(0.455)	[0.001]	[0.001]
Female student	0.585	0.552	0.033***	-0.001	0.555	0.554	0.001	-0.003***
	(0.493)	(0.497)	[0.001]	[0.001]	(0.497)	(0.497)	[0.001]	[0.001]
Age of student at test	23.33	23.45	-0.129***	-0.146***	23.21	23.47	-0.256***	-0.165***
	(8.999)	(9.245)	[0.022]	[0.023]	(9.194)	(9.231)	[0.021]	[0.021]
Number of driving tests	2.647	2.350	0.297***	0.057***	2.387	2.368	0.019***	0.022***
	(2.092)	(1.824)	[0.005]	[0.005]	(1.878)	(1.842)	[0.004]	[0.004]
Number of theory tests	2.852	2.437	0.415***	0.083***	2.502	2.462	0.040***	0.048***
	(3.183)	(2.771)	[0.008]	[0.008]	(2.898)	(2.795)	[0.007]	[0.007]

Notes. Standard deviations are in parentheses in columns 1-2 and 5-6. Standard errors are in brackets in columns 3-4 and 7-8. Each entry in columns 3 and 4 is derived from a separate OLS regression where the explanatory variable is an indicator for Arab tester. Each entry in columns 7 and 8 is derived from a separate OLS regression where the explanatory variable is an indicator for Female tester. Columns 4 and 8 include test area fixed effects.

*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Table 4
Ethnic Bias

	Dependent Variable: Test Outcome (Pass=1)				
	(1)	(2)	(3)	(4)	(5)
Arab student	-0.055***	-0.055***	-0.034***	-0.034***	-0.034***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Arab tester	-0.048***	-0.036**	-0.035**	-0.018	
	(0.017)	(0.016)	(0.016)	(0.017)	
Arab student x Arab tester	0.071***	0.066***	0.065***	0.063***	0.059***
	(0.011)	(0.012)	(0.011)	(0.013)	(0.013)
Test area fixed effects	Yes	Yes	Yes	Yes	Yes
Time controls	No	Yes	Yes	Yes	Yes
Student characteristics	No	No	Yes	Yes	Yes
Tester characteristics	No	No	No	Yes	Yes
Tester fixed effects	No	No	No	No	Yes
Observations	2,615,921	2,615,921	2,615,921	2,615,921	2,615,921
R-squared	0.021	0.024	0.036	0.048	0.074

Notes. Time controls include fixed effects for test year, month and day of week. Student characteristics include a female indicator, age (divided by 100), current driving test number and number of theory tests. Tester characteristics include a female indicator (columns 1-4), age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Table 5
Gender Bias

	Dependent Variable: Test Outcome (Pass=1)				
	(1)	(2)	(3)	(4)	(5)
Female student	-0.075***	-0.076***	-0.076***	-0.075***	-0.075***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Female tester	0.002	0.007	0.007	0.025	
	(0.030)	(0.030)	(0.030)	(0.029)	
Female student x Female tester	-0.044***	-0.045***	-0.045***	-0.046***	-0.042***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.012)
Test area fixed effects	Yes	Yes	Yes	Yes	Yes
Time controls	No	Yes	Yes	Yes	Yes
Student characteristics	No	No	Yes	Yes	Yes
Tester characteristics	No	No	No	Yes	Yes
Tester fixed effects	No	No	No	No	Yes
Observations	2,615,921	2,615,921	2,615,921	2,615,921	2,615,921
R-squared	0.026	0.029	0.036	0.048	0.074

Notes. Time controls include fixed effects for test year, month and day of week. Student characteristics include an Arab indicator, age (divided by 100), current driving test number and number of theory tests. Tester characteristics include an Arab indicator (columns 1-4), age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Table 6
Ethnic and Gender Biases

	Dependent Variable: Test Outcome (Pass=1)				
	(1)	(2)	(3)	(4)	(5)
Arab student x Arab tester	0.069***	0.064***	0.064***	0.063***	0.059***
	(0.012)	(0.012)	(0.011)	(0.013)	(0.013)
Female student x Female tester	-0.045***	-0.045***	-0.045***	-0.046***	-0.042***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.012)
Test area fixed effects	Yes	Yes	Yes	Yes	Yes
Time controls	No	Yes	Yes	Yes	Yes
Student characteristics	No	No	Yes	Yes	Yes
Tester characteristics	No	No	No	Yes	Yes
Tester fixed effects	No	No	No	No	Yes
Observations	2,615,921	2,615,921	2,615,921	2,615,921	2,615,921
R-squared	0.027	0.030	0.036	0.048	0.074

Notes. Time controls include fixed effects for test year, month and day of week. Student characteristics include a female indicator, an Arab indicator, age (divided by 100), current driving test number and number of theory tests. Tester characteristics include a female indicator (columns 1-4), an Arab indicator (columns 1-4), age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Table 7
Tester or Student Behavior?

	Dependent Variable: Test Outcome (Pass=1)				
	(1)	(2)	(3)	(4)	(5)
Arab student	-0.059***	-0.059***	-0.037***	-0.037***	-0.037***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Tester from integrated locality	0.025	0.023	0.022	0.018	
	(0.020)	(0.018)	(0.018)	(0.015)	
Tester from integrated locality x Arab student	0.031***	0.029***	0.029***	0.024**	0.024**
	(0.011)	(0.011)	(0.011)	(0.010)	(0.010)
Test area fixed effects	Yes	Yes	Yes	Yes	Yes
Time controls	No	Yes	Yes	Yes	Yes
Student characteristics	No	No	Yes	Yes	Yes
Tester characteristics	No	No	No	Yes	Yes
Tester fixed effects	No	No	No	No	Yes
Observations	2,436,935	2,436,935	2,436,935	2,436,935	2,436,935
R-squared	0.022	0.025	0.037	0.049	0.076

Notes. The analysis is restricted to Jewish testers. "Tester from an integrated locality" is an indicator that equals 1 if the tester resides in an ethnically mixed locality and 0 otherwise (i.e. the tester resides in a Jewish locality). Time controls include fixed effects for test year, month and day of week. Student characteristics include a female indicator, age (divided by 100), current driving test number and number of theory tests. Tester characteristics include a female indicator (columns 1-4), age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Table 8
Ethnic and Gender Biases, with student Fixed Effects

	Dependent Variable: Test Outcome (Pass=1)			
	(1)	(2)	(3)	(4)
Arab student x Arab tester	0.059***	0.042***		0.041***
	(0.013)	(0.011)		(0.011)
Female student x Female tester	-0.042***		-0.046***	-0.046***
	(0.012)		(0.011)	(0.011)
Test area fixed effects	Yes	Yes	Yes	Yes
Time controls	Yes	Yes	Yes	Yes
Student characteristics	Yes	Yes	Yes	Yes
Tester characteristics	Yes	Yes	Yes	Yes
Tester fixed effects	Yes	Yes	Yes	Yes
Student fixed effects	No	Yes	Yes	Yes
Observations	2,615,921	2,615,921	2,615,921	2,615,921
Identifying observations	2,615,921	482,012	575,402	877,594
R-squared	0.074	0.583	0.583	0.583

Notes. Time controls include fixed effects for test year, month and day of week. Student characteristics include a female indicator, an Arab indicator, age (divided by 100), current driving test number and number of theory tests. Tester characteristics include age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Table 9
The Effect of Tester Experience on Bias

	Dependent Variable: Test Outcome (Pass=1)			
	(1)	(2)	(3)	(4)
Arab student x Arab tester	0.059***	0.018	0.060***	0.020
	(0.013)	(0.131)	(0.013)	(0.131)
Female student x Female tester	-0.042***	-0.042***	-0.042	-0.041
	(0.012)	(0.012)	(0.090)	(0.090)
Arab student x Arab tester x Tester age		0.087		0.083
		(0.248)		(0.248)
Female student x Female tester x Tester age			0.019	0.017
			(0.182)	(0.182)
Additional interactions	No	Yes	Yes	Yes
Test area fixed effects	Yes	Yes	Yes	Yes
Time controls	Yes	Yes	Yes	Yes
Student characteristics	Yes	Yes	Yes	Yes
Tester characteristics	Yes	Yes	Yes	Yes
Tester fixed effects	Yes	Yes	Yes	Yes
Observations	2,615,921	2,615,921	2,615,921	2,615,921
R-squared	0.074	0.074	0.075	0.075

Notes. Additional interactions vary across columns: in column 2 they include interactions between tester age and indicators for Arab student and Arab tester; in column 3 they include interactions between tester age and indicators for female student and female tester; in column 4 they include both sets of interactions. Time controls include fixed effects for test year, month and day of week. Student characteristics include a female indicator, an Arab indicator, age (divided by 100), current driving test number and number of theory tests. Tester characteristics include age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Table 10
The Effect of Physical Distance

	Dependent Variable: Test Outcome (Pass=1)		
		Restricted sample	
	Private vehicle tests	Private vehicle tests	Motorcycle tests
	(1)	(2)	(3)
Arab student	-0.034***	-0.038***	-0.021**
	(0.004)	(0.007)	(0.008)
Arab student x Arab tester	0.059***	0.045***	-0.018
	(0.013)	(0.008)	(0.019)
Test area fixed effects	Yes	Yes	Yes
Time controls	Yes	Yes	Yes
Student characteristics	Yes	Yes	Yes
Tester characteristics	Yes	Yes	Yes
Tester fixed effects	Yes	Yes	Yes
Observations	2,615,921	961,760	282,871
R-squared	0.074	0.065	0.108

Notes. The analysis in columns 2 and 3 is restricted to testers who conduct both private vehicle and motorcycle tests. Time controls include fixed effects for test year, month and day of week. Student characteristics include a female indicator, age (divided by 100), current driving test number and number of theory tests. Tester characteristics include age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Table 11
The Effect of Temperature on Bias

	Dependent Variable: Test Outcome (Pass=1)			
	Jewish Testers	Arab Testers	Male Testers	Female Testers
	(1)	(2)	(3)	(4)
Arab student	-0.020***	0.017		
	(0.006)	(0.020)		
Maximum temperature	-0.026***	-0.041		
	(0.012)	(0.048)		
Maximum temperature x Arab student	-0.054***	0.003		
	(0.015)	(0.036)		
Female student			-0.074***	-0.125***
			(0.008)	(0.016)
Maximum temperature			-0.047***	-0.024
			(0.015)	(0.048)
Maximum temperature x Female student			-0.003	0.041
			(0.016)	(0.050)
Test area fixed effects	Yes	Yes	Yes	Yes
Time controls	Yes	Yes	Yes	Yes
Student characteristics	Yes	Yes	Yes	Yes
Tester characteristics	Yes	Yes	Yes	Yes
Tester fixed effects	Yes	Yes	Yes	Yes
Observations	2,330,579	169,619	2,297,760	202,438
R-squared	0.075	0.040	0.071	0.095

Notes. Maximum temperature is the highest temperature measured on the day of the test in the weather station closest to the test center. Time controls include fixed effects for test year, month and day of week. Student characteristics include a female indicator (columns 1 and 2), an Arab indicator (columns 3 and 4), age (divided by 100), current driving test number and number of theory tests. Tester characteristics include age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Figure 1:

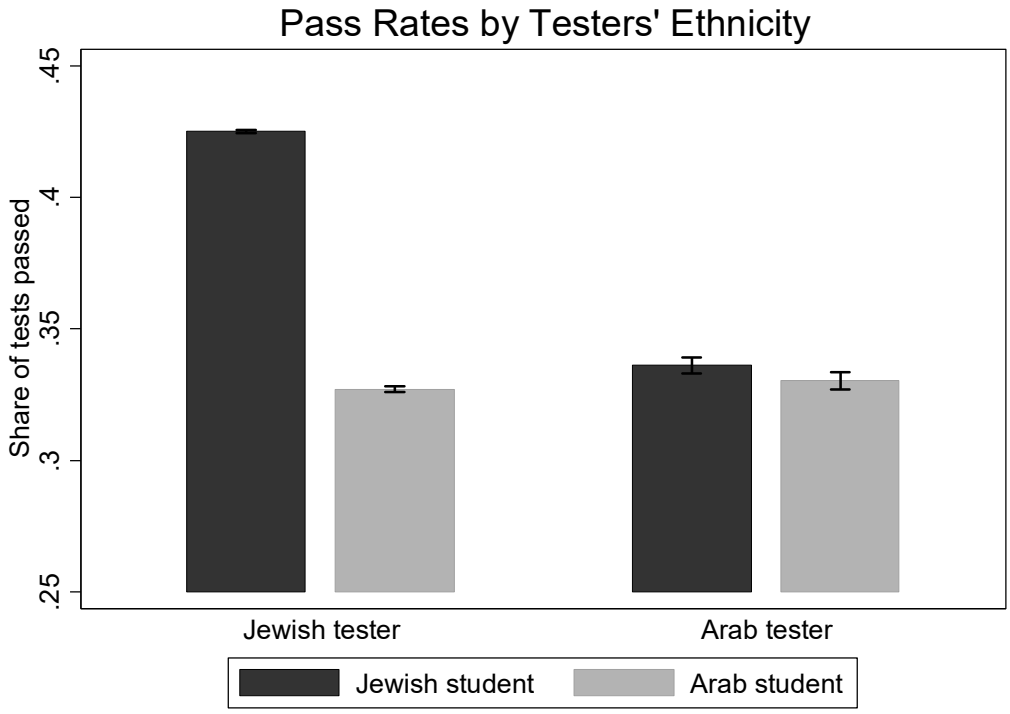


Figure 2:

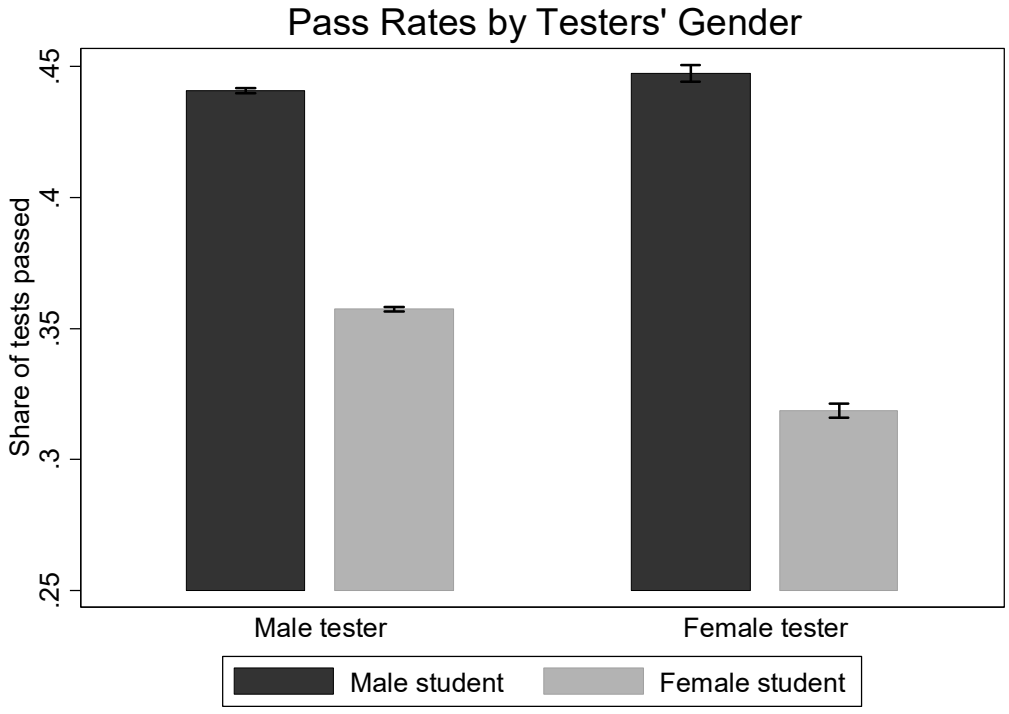


Figure 3:

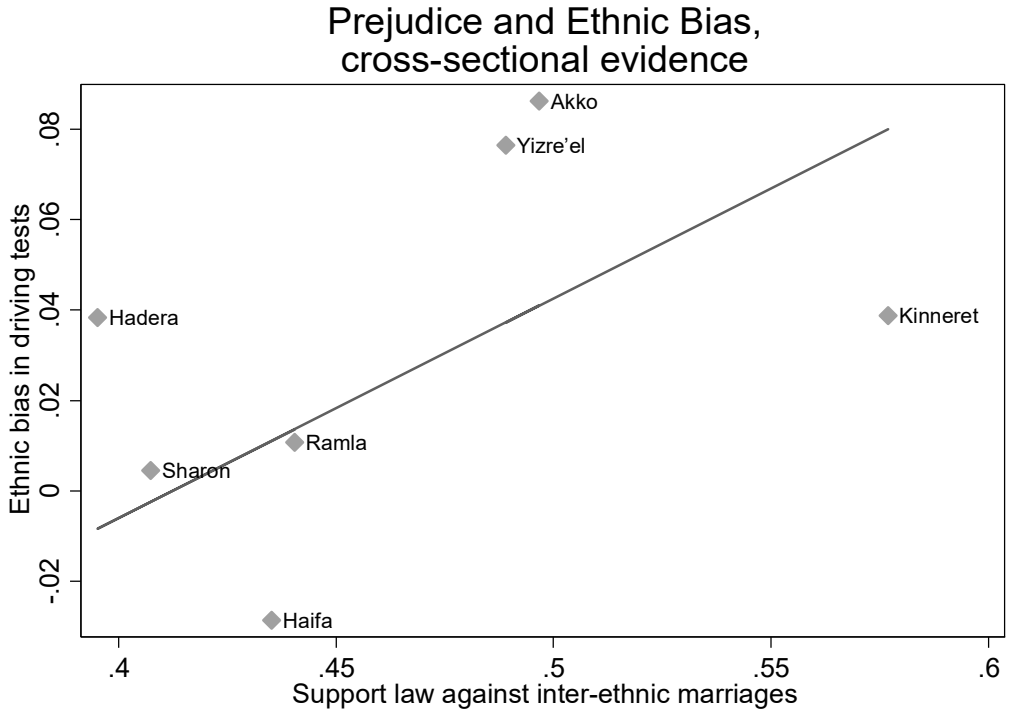


Figure 4:

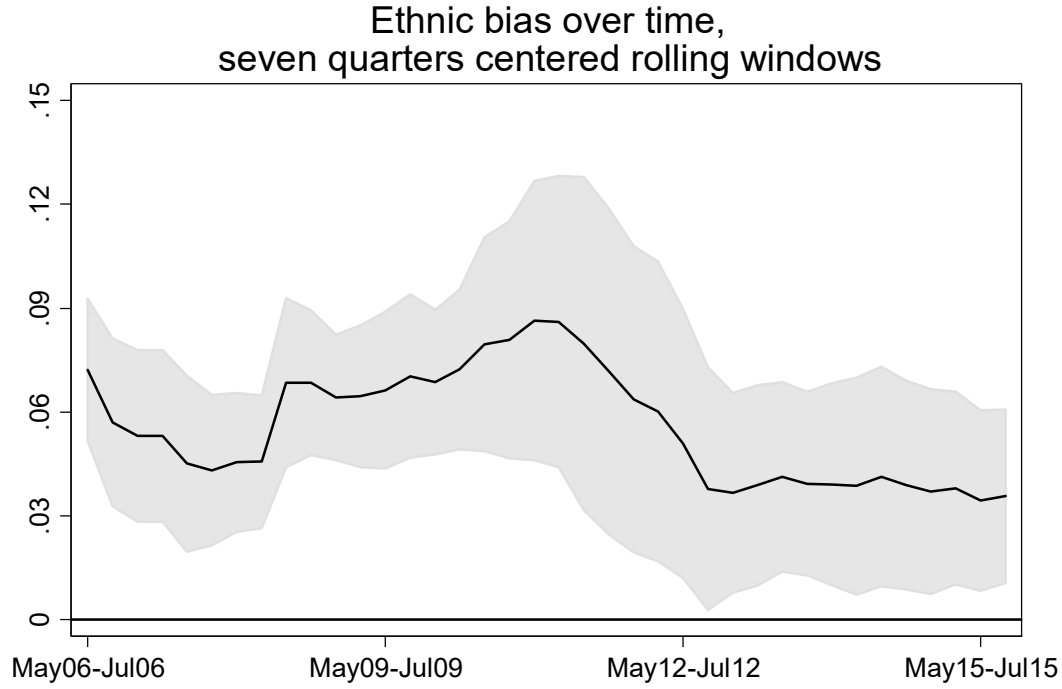
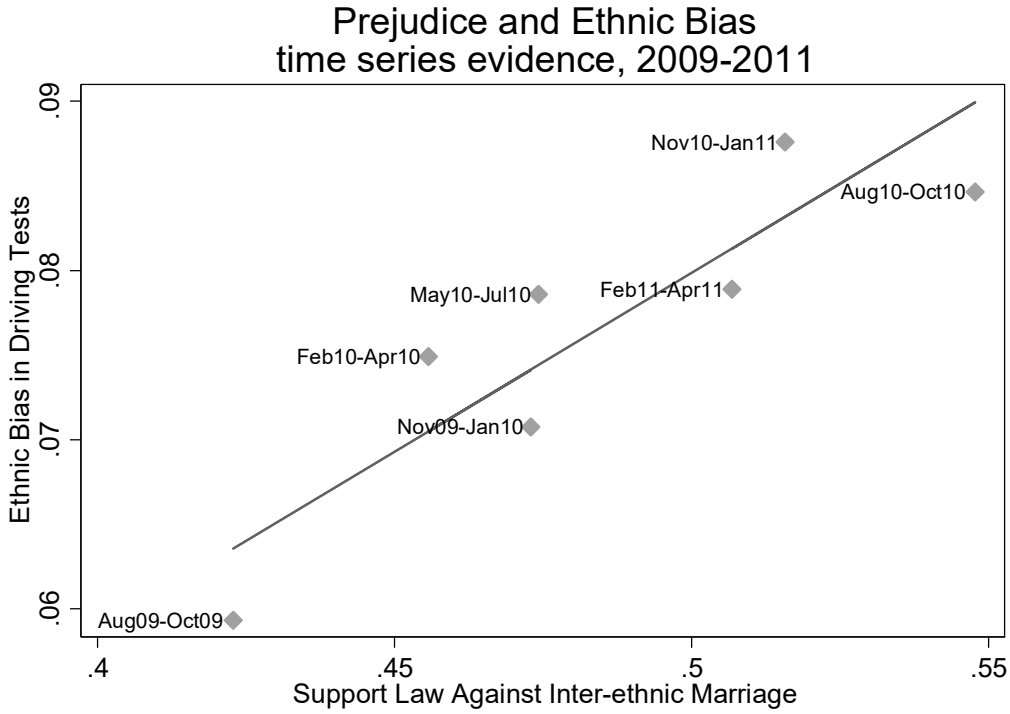


Figure 5:


Prejudice and Ethnic Bias
time series evidence, 2009-2011
Ethnic Bias in Driving Tests
.06
.07
.08
.09
Support Law Against Inter-ethnic Marriage
.4
.45
.5
.55
Aug09-Oct09
Nov09-Jan10
Feb10-Apr10
May10-Jul10
Feb11-Apr11
Aug10-Oct10
Nov10-Jan11

Appendix A

Appendix Table A1
Shares of Tests Passed, by Ethnicity of Student and Tester

	Arab student	Jewish student	Difference
	(1)	(2)	(3)
Arab tester	0.330	0.336	-0.060***
	(0.470)	(0.472)	[0.002]
	N=81,986	N=97,000	N=178,986
Jewish tester	0.327	0.425	-0.098***
	(0.469)	(0.494)	[0.001]
	N=686,537	N=1,750,398	N=2,436,935
Difference	0.003*	-0.089***	0.092***
	[0.002]	[0.002]	[0.002]
	N=768,523	N=1,847,398	N=2,615,921

Notes. Standard deviations in parentheses and standard errors in brackets.
Column 3 and row 3 are estimated using OLS.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Appendix Table A2
Shares of Tests Passed, by Gender of Student and Tester

	Female student	Male student	Difference
	(1)	(2)	(3)
Female tester	0.319	0.447	-0.129***
	(0.466)	(0.497)	[0.002]
	N=116,568	N=93,295	N=209,863
Male tester	0.357	0.441	-0.083***
	(0.479)	(0.497)	[0.001]
	N=1,334,132	N=1,071,926	N=2,406,058
Difference	-0.039***	0.007***	-0.045***
	[0.002]	[0.002]	[0.002]
	N=1,450,700	N=1,165,221	N=2,615,921

Notes. Standard deviations in parentheses and standard errors in brackets.
Column 3 and row 3 are estimated using OLS.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Appendix Table A3
Ethnic Bias and Other Tester Characteristics

	Dependent Variable: Test Outcome (Pass=1)				
	(1)	(2)	(3)	(4)	(5)
Arab student x Arab tester	0.059***	0.059***	0.063***	0.060***	0.063***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Arab student x Female tester		-0.003			0.001
		(0.010)			(0.010)
Arab student x Tester age			0.077*		0.079*
			(0.047)		(0.047)
Arab student x Number of same day tests				-0.001	-0.001
				(0.001)	(0.001)
Test area fixed effects	Yes	Yes	Yes	Yes	Yes
Time controls	Yes	Yes	Yes	Yes	Yes
Student characteristics	Yes	Yes	Yes	Yes	Yes
Tester characteristics	Yes	Yes	Yes	Yes	Yes
Tester fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	2,615,921	2,615,921	2,615,921	2,615,921	2,615,921
R-squared	0.074	0.074	0.074	0.074	0.074

Notes. Time controls include fixed effects for test year, month and day of week. Student characteristics include an Arab indicator, a female indicator, age (divided by 100), current driving test number and number of theory tests. Tester characteristics include age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Appendix Table A4
Gender Bias and Other Tester Characteristics

	Dependent Variable: Test Outcome (Pass=1)				
	(1)	(2)	(3)	(4)	(5)
Female student x Female tester	-0.042***	-0.042***	-0.033***	-0.042***	-0.031**
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Female student x Arab tester		0.035			0.048*
		(0.026)			(0.026)
Female student x Tester age			0.197***		0.226***
			(0.069)		(0.070)
Female student x Number of tests on each day				0.001	0.000
				(0.000)	(0.000)
Test area fixed effects	Yes	Yes	Yes	Yes	Yes
Time controls	Yes	Yes	Yes	Yes	Yes
Student characteristics	Yes	Yes	Yes	Yes	Yes
Tester characteristics	Yes	Yes	Yes	Yes	Yes
Tester fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	2,615,921	2,615,921	2,615,921	2,615,921	2,615,921
R-squared	0.074	0.074	0.074	0.074	0.074

Notes. Time controls include fixed effects for test year, month and day of week. Student characteristics include an Arab indicator, a female indicator, age (divided by 100), current driving test number and number of theory tests. Tester characteristics include age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Appendix Table A5
Ethnic Bias – Different Methods for Identifying Ethnicity

	Dependent Variable: Test Outcome (Pass=1)				
	(1)	(2)	(3)	(4)	(5)
Arab student	-0.034***	-0.035***	-0.044***	-0.034***	-0.044***
	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)
Arab student x Arab tester	0.059***	0.054***	0.059***	0.052***	0.066***
	(0.013)	(0.014)	(0.014)	(0.013)	(0.014)
Test area fixed effects	Yes	Yes	Yes	Yes	Yes
Time controls	Yes	Yes	Yes	Yes	Yes
Student characteristics	Yes	Yes	Yes	Yes	Yes
Tester characteristics	Yes	Yes	Yes	Yes	Yes
Tester fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	2,615,921	2,613,309	2,615,381	2,615,921	2,615,381
R-squared	0.074	0.074	0.074	0.074	0.074

Notes. In column 1 we replicate the analysis of ethnic bias using the original ethnicity classification (column 5 of Table 4). In column 2 we identify a name as Arab if it is at least three times more popular among Arabs than it is among Jews, and as Jewish if it is at least three times more popular among Jews than it is among Arabs. In column 3 we identify student ethnicity first by locality of residence and then by name, and tester ethnicity first by name and then by locality of residence. In column 4 we identify student ethnicity first by name and then by locality of residence, and tester ethnicity first by locality of residence and then by name. In column 5 we identify both student ethnicity and tester ethnicity first by locality of residence and then by name. Time controls include fixed effects for test year, month and day of week. Student characteristics include a female indicator, age (divided by 100), current driving test number and number of theory tests. Tester characteristics include age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Appendix Table A6
Ethnic and Gender Biases, with Driving Teacher FE

	Dependent Variable: Test Outcome (Pass=1)	
	(1)	(2)
Arab student x Arab tester	0.059***	0.060***
	(0.013)	(0.012)
Female student x Female tester	-0.043***	-0.043***
	(0.012)	(0.012)
Test area fixed effects	Yes	Yes
Time controls	Yes	Yes
Student characteristics	Yes	Yes
Tester characteristics	Yes	Yes
Tester fixed effects	Yes	Yes
Driving teacher fixed effects	No	Yes
Observations	2,527,832	2,527,832
R-squared	0.073	0.098

Notes. The analysis in this table is restricted to students for whom we have a driving teacher identifier. Time controls include fixed effects for test year, month and day of week. Student characteristics include an Arab indicator, a female indicator, age (divided by 100), current driving test number and number of theory tests. Tester characteristics include age (divided by 100) and total number of same day tests.
Estimated using OLS. Standard errors, clustered by tester, are in parentheses.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Appendix B

Appendix Table B1
Geographical Distribution of Motorcycle Driving Tests, by MOT Regions

MOT Region	Number of Test Areas	Tester: Student:	Jewish Jewish	Jewish Arab	Arab Jewish	Arab Arab	Tests
Tel Aviv and Center	14		89.11	10.82	0.05	0.02	150,978
Haifa and North	14		56.66	27.65	10.13	5.56	47,254
Be'er Sheba and the Negev	10		90.41	9.58	0.01	0.00	23,876
Jerusalem and South	5		74.23	25.12	0.51	0.15	60,763
Countrywide	43		80.60	16.60	1.83	0.97	282,871

Notes. The table shows, for each MOT region, the share (in %) of driving tests in each combination of student and tester ethnicities.

Appendix Table B2
Summary Statistics for Motorcycle Tests

	Panel A: Students (N=180,002)			
	All students	Arab students	Jewish students	Diff
	(1)	(2)	(3)	(4)
Arab student	0.166 (0.372)	1 (0.000)	0 (0.000)	1 [N/A]
Female student	0.091 (0.287)	0.024 (0.152)	0.104 (0.305)	-0.081*** [0.001]
Age in test	26.46 (9.089)	26.19 (8.359)	26.52 (9.225)	-0.323*** [0.054]
Number of driving tests	1.286 (0.465)	1.333 (0.500)	1.277 (0.457)	0.056*** [0.003]
Number of theory tests	0.552 (1.361)	0.594 (1.604)	0.543 (1.307)	0.051*** [0.010]

Notes. Standard deviations are in parentheses in columns 1-3. Standard errors are in brackets in column 4. Each entry in column 4 is derived from a separate OLS regression where the explanatory variable is an indicator for Arab student. Number of driving tests is the current test number, i.e. number of previous failed tests plus one. Number of theory tests is the number of theory tests the student has taken.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Appendix Table B2
Summary Statistics for Motorcycle Tests

	Panel B: Testers (N=70)			
	All testers	Arab testers	Jewish testers	Diff
	(1)	(2)	(3)	(4)
Arab tester	0.043 (0.204)	1 (0.000)	0 (0.000)	1 [N/A]
Female tester	0.014 (0.120)	0.000 (N/A)	0.015 (0.122)	-0.015 [0.015]
Age in test	53.03 (6.989)	46.65 (2.991)	53.32 (6.990)	-6.673*** [1.669]
Number of same day tests	17.75 (6.272)	19.78 (0.815)	17.66 (6.397)	2.121** [0.878]

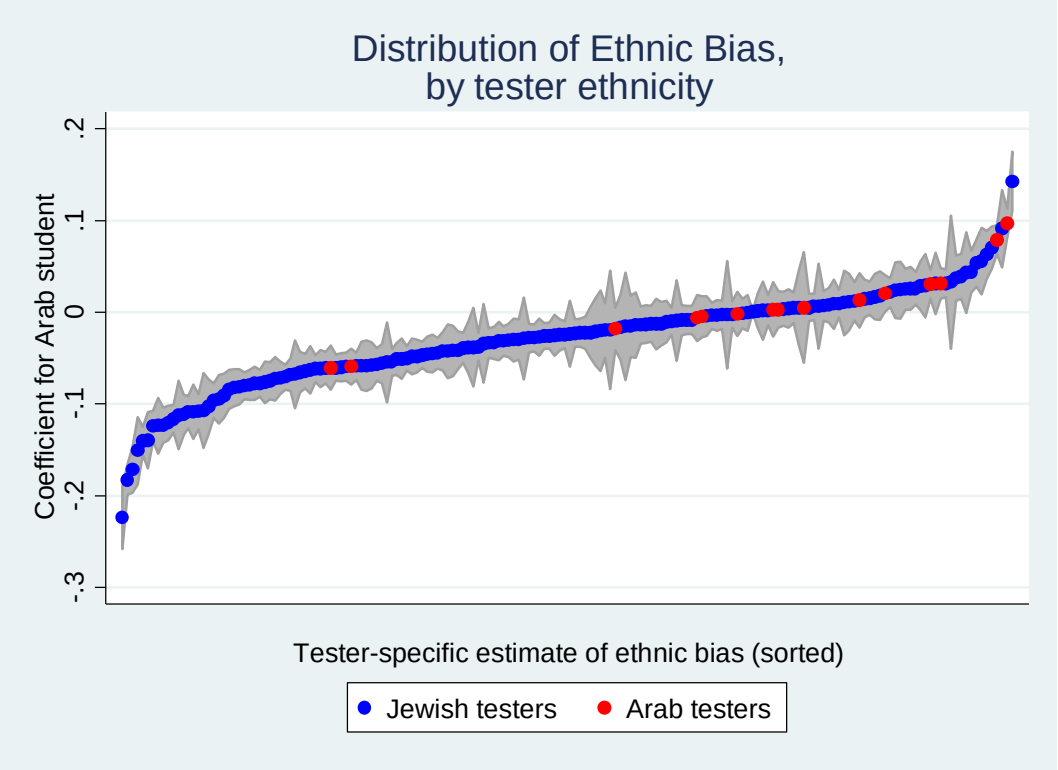
Notes. Standard deviations are in parentheses in columns 1-3. Standard errors are in brackets in column 4. Each entry in column 4 is derived from a separate OLS regression where the explanatory variable is an indicator for Arab tester. Number of same day tests is the total number of tests the tester conducted on the day of the observed test.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Appendix Table B3
Balancing Tests for the Assignment of Students to Testers in Motorcycle Tests, By Ethnicity

	Mean		Differences in Means Arab vs. Jewish Tester	
	Arab Tester	Jewish Tester	No controls	w/ Area FE
	(1)	(2)	(3)	(4)
Arab student	0.347	0.171	0.176***	0.015***
	(0.476)	(0.376)	[0.005]	[0.005]
Female student	0.034	0.078	-0.045***	-0.009***
	(0.180)	(0.269)	[0.002]	[0.002]
Age of student at test	25.84	26.07	-0.232**	-0.559***
	(9.499)	(9.150)	[0.108]	[0.117]
Number of driving tests	1.674	1.557	0.117***	0.114***
	(0.994)	(0.922)	[0.011]	[0.012]
Number of theory tests	0.608	0.634	-0.026**	-0.046***
	(1.400)	(1.462)	[0.016]	[0.018]

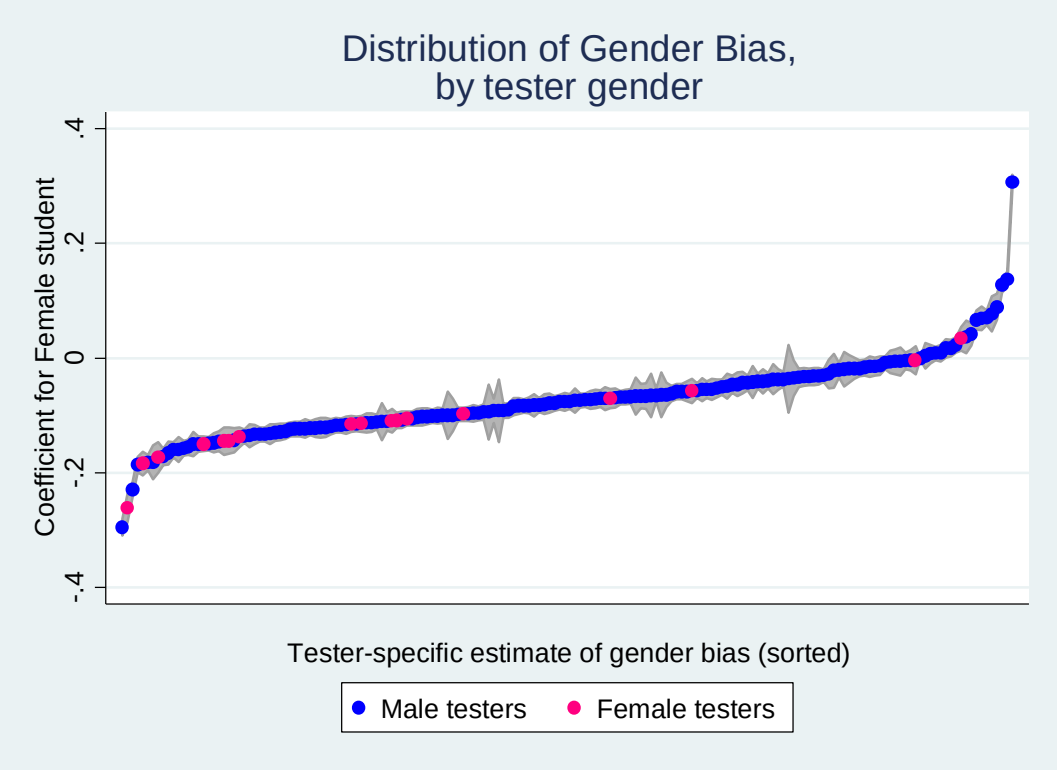
Notes. Standard deviations are in parentheses in columns 1-2. Standard errors are in brackets in columns 3-4. Each entry in columns 3 and 4 is derived from a separate OLS regression where the explanatory variable is an indicator for Arab tester. Column 4 includes test area fixed effects.
*, **, *** represent statistical significance at the 10%, 5%, and 1% levels.

Appendix Figure 1A:



Notes: The figure plots tester-specific estimates of ethnic bias together with 95% confidence intervals. The estimates are derived from regressions – run separately for each tester – of test outcome on an Arab student indicator, time controls (fixed effects for test year, month and day of week); student characteristics (a female indicator, age, current test number and number of theory tests); and tester characteristics (age and total number of same day tests). The figure reports estimates for testers who conducted at least 1,000 tests.

Appendix Figure 1B:



Notes: The figure plots tester-specific estimates of gender bias together with 95% confidence intervals. The estimates are derived from regressions – run separately for each tester – of test outcome on a female student indicator, time controls (fixed effects for test year, month and day of week); student characteristics (an Arab indicator, age, current test number and number of theory tests); and tester characteristics (age and total number of same day tests). The figure reports estimates for testers who conducted at least 1,000 tests.