Racial Bias and In-Group Bias in Virtual Reality Courtrooms

Non Peer-reviewed author version

BIELEN, Samantha; MARNEFFE, Wim & MOCAN, Naci (2021) Racial Bias and In-Group Bias in Virtual Reality Courtrooms. In: The Journal of Law and Economics, 64 (2) , p. 269 -300.

DOI: 10.1086/712421
Handle: http://hdl.handle.net/1942/35551

Racial Bias and In-group Bias in Judicial Decisions: Evidence from Virtual Reality Courtrooms

Samantha Bielen
Hasselt University

Wim Marneffe
Hasselt University

Naci Mocan
Louisiana State University, NBER and IZA

November 2018

We shot videos of criminal trials using 3D Virtual Reality (VR) technology, prosecuted by actual prosecutors and defended by actual defense attorneys in an actual courtroom. This is the first paper that utilizes VR technology in a non-computer animated setting, which allowed us to replace white defendants in the courtroom with individuals who have Middle Eastern or North African descent in a real-life environment. This enabled us to alter only the race of the defendants in these trials, holding all activity in the courtroom constant (http://proficient.ninja/splitscreen/). Law students, economics students and practicing lawyers are randomly assigned to watch with VR headsets, from the view point of the judge, these trials that differed only in defendants' skin color. Background information obtained from the evaluators allowed us to identify their cultural heritage. In this experiment evaluators made decisions on guilt/innocence in these burglary and assault cases, as well as prison sentence length and fine in accordance with the guidelines provided by the relevant law. The results reveal bias in conviction decisions against minorities, emanated from both white and minority evaluators. We find in-group favoritism in prison sentences and fines, driven by white evaluators. This translates into overall racial bias against minority defendants in prison sentences and fines. We find only scant evidence that the concerns of the evaluators about terrorism, about immigration, or their trust in the judiciary or the police have an impact on their judicial decisions, suggesting that the source of the bias may be deep-rooted.

We thank Christine Jolls, Ozkan Eren, Shawn Kantor, Carl Kitchens, Marc Isaac, Bulent Unel, Stefan Voigt, Barton Willage, Richard Boylan and participants of seminars at the University of Iowa, Florida State University, the European Law and Economics Association Meetings in Milan, Italy, and the Economic Consequences of Legal Development Workshop at UCLouvain for helpful comments. Funding from the Faculty of Business Economics of Hasselt University is gratefully acknowledged. We are grateful to Johan Ackaert and Sofie Hennau for allowing us to engage their law students in the study, and to the Office of the District Attorney of Limburg, Belgium for sharing their expertise and for allowing us to use their main courtroom. We are grateful to Jeff Luyten and Michel Séquaris from *Uncanny Virtual Experiences* for making this project possible by providing technical support in creating the VR videos and to Sem Metten for creating the website that communicated with the VR videos and that enabled the participants to register their decisions. Bielen acknowledges financial support from Research Foundation Flanders, grant number 12S3117N.

Racial Bias and In-group Bias in Judicial Decisions:
Evidence from Virtual Reality Courtrooms

I. Introduction

The investigation of racial bias in decision-making is important for both scientific inquiry and for public policy. A particularly important decision is the judgment passed on others with consequential outcomes. For example, although judicial decisions are expected to be made blindly, whether such impartiality exists in practice has long been debated, and a great deal of research has focused on the investigation of whether conviction, sentencing and clemency decisions made by judges, juries or state governors are biased against minorities (Alesina and La Ferrara 2014, Rehavi and Starr 2014, Abrams, Bertrand and Mullainathan 2012, Argys and Mocan 2004, Glaeser and Sacerdote 2003, Mustard 2001). These findings are important because judicial decisions are made with deliberation (as opposed to quickly), and they are high-stakes decisions with serious consequences.[1]

A related, and more nuanced line of research has recently emerged to investigate the existence of in-group bias (decision-maker's preferential treatment of the members of his/her own group). Anwar, Bayer and Hjalmarsson (2012) analyzed the impact of racial composition of juries in Florida, and found that all-white juries are more likely to convict black defendants. Shayo and Zussman (2011) analyzed data from Israeli small claims courts and reported that a claim was more likely to be accepted by the judge if the judge is of the same ethnicity as the plaintiff, and that this effect was strongly associated with terrorism intensity in the area surrounding the court during the year preceding the judicial decision. Gazal-Ayal and Sulitzeanu-Kenan (2010) found in-group bias in Jewish and Arab judges' decisions on criminal cases in Israel. Grossman et al. (2015) reported that Arab defendants received more lenient

[1] Racial disparities in decisions that are made quickly (almost by reflex or intuition) have been detected in the calls of basketball referees (Price and Wolfers 2010) and baseball umpires (Parsons et al. 2011).

punishment in the appellate court when the panel of judges included at least one Arab judge. Depew, Eren and Mocan (2017) reported negative in-group bias in judicial decisions in Louisiana juvenile courts.

Although recent research exploited random assignment of defendants to judges to alleviate some of the inherent selection issues associated with the analysis of judge decisions (Eren and Mocan 2018, Shayo and Zussman 2017, Depew, Eren and Mocan 2017, Abrams, Bertrand and Mullainathan 2012, Shayo and Zussman 2011), random assignment does not resolve all selection issues. Even if the defendants are randomly assigned to judges, there are a number of intervening steps before the defendant comes in front of the judge, and the events during this process may confound the inference obtained from the analysis of judicial decisions.[2]

Similarly, the composition of cases that come in front of a judge can also change because of the strategic behavior of defense attorneys. For example, after the case is assigned to a judge the defense attorney may be more or less open to plea bargaining based on the attributes of the judge, such as the race and the reputation of the judge.

The impression of judicial racial bias may also emerge as an artifact of the behavior of the individuals in the courtroom. For example, if prosecutors are more diligent and aggressive towards a certain type of defendant in comparison to others, this differential effort in the courtroom could have a systematic effect on judge decisions. Similarly, the effort and effectiveness of the defense attorneys may systematically differ between groups of defendants. For instance, minority defendants may not be able to afford high-quality attorneys if their

[2] For instance, after cases are (randomly) assigned to judges, prosecutors enter the process and they have the discretion to determine the charges levied against the defendants. The attitude of the prosecutor regarding charging a particular defendant more or less aggressively has an impact on the eventual outcome of the case. For example, as described in Boylan (2012), a prosecutor can charge a defendant with possessing a gun when selling drugs in addition to a drug-trafficking offense, while deciding not to file a gun charge for another drug seller. If prosecutors are more lenient towards a group of defendants on the margin, the conviction and sentencing decisions of judges will resemble favoritism towards that group of defendants even if judges themselves are completely unbiased. Furthermore, the willingness of the prosecutor to offer a plea bargain arrangement to the defendant and his/her discretion in setting the terms of the offer have an impact on which defendants and what types of cases end up being adjudicated.

income and wealth levels are lower, and ineffective defense provided by low quality attorneys will translate into worse outcomes for minorities such as higher conviction probabilities and longer sentences (Anderson and Heaton 2012). Part of the negative outcomes associated with court-assigned attorneys to low-income defendants is attributable to attorney effort, motivated by the compensation structure of these attorneys (Agan et al. 2018). Thus, racial differences in judicial decisions could be driven by attorney quality and effort, not by the racial bias of judges.[3]

Finally, particular circumstances of each case are different, and each particular case that is brought to trial has its own set of mitigating and aggravating circumstances, its own features about the background and the criminal history of the defendant, and the circumstances of the victim (if the crime involves a victim). Researchers try to account for such differences by controlling for the type of crime that is allegedly committed by the defendant (e.g. burglary, assault and so on). Such broad classification of crime commission may be sufficient if the number of cases handled by each judge is large, but this may not always be the case in practice, or in a particular data set analyzed.[4] These pre-trial decisions and the selection problems they create, as well as the actions of various parties in the courtroom (ranging from the intensity of the language used by both the defense and the prosecutor to the body language), are nearly impossible to control for.

In this paper we provide a solution to the biases caused by these confounders. Using Virtual Reality (VR) technology we record trials in a real courtroom, with actual prosecutors and actual

[3] Chen, Halberstam and Yu (2018) show that voice-based snap judgements of the evaluators (who made assessments about the masculinity of the voices) by listening to lawyers' identical short opening remarks in front out the U.S. Supreme Court, predict the outcomes in the Court.

[4] Other, more subtle, complications may also emerge in the investigation of in-group bias. For example, assume that within a crime category minority defendants are more likely to have A-type circumstances (e.g. stealing a car and getting involved in an accident using that car) and that white defendants are more likely to have B-type circumstances (e.g. stealing a car and committing a robbery using that car). Assume that judges are not racially biased, but minority judges are more strict against B-type circumstance and that white judges are more strict against A-type circumstance. In this scenario, if judges have discretion in sentencing, white (minority) judges will be harsher against minority (white) defendants, not because of racial bias but because of differential attitudes of judges towards different offenses.

defense attorneys. This is the first paper that utilizes 3D Virtual Reality technology that uses actual people in these videos, rather than computer-animated scenes. The details are provided in Sections II and III.[5]

The VR technology allows us to replace the defendants in the 3D VR videos of the trials, holding constant every spoken word and every action in the court, enabling us to create arguably perfect counterfactuals. A glimpse of one of the trials with two different defendants can be seen at http://proficient.ninja/splitscreen/.

One-hundred ninety-two evaluators, who are either Master's level law students, undergraduate economics students, or practicing criminal attorneys, made decisions on conviction as well as on prison sentence length and fine (on those defendants who are found guilty) in accordance with the guidelines provided by the relevant penal code. By design, the race of the defendant is uncorrelated with the characteristics of both the prosecutors and the defense attorneys and with all courtroom activity. Defendant race is also uncorrelated with evaluator attributes. This feature of the design allows us to decompose the sources of biases in decision making; i.e. it enables us to investigate whether the bias is driven by white or minority (or both) evaluators.

It should be underscored that the entire variation within a given trial is obtained from the skin color of the defendants. We hold the prosecutor, the defense attorney and their actions in the courtroom constant, and we do not reveal the names of the defendants to the evaluators. The defendants speak only three times during the trial to answer three questions posed by the judge. These answers are: "Yes," "Not guilty", and "No". All actor-defendants are born in Belgium

[5] Schulman et al. (1999) shot 2D videos of actors who portrayed patients displaying various symptoms of cardiac problems. These videos were shown to physicians and physicians' estimates of the probability of coronary artery disease and their recommendations about cardiac catheterization are recorded to analyze differences in recommended treatment by race, sex, and age of the patient. Although this is a creative and important study, the 144 treatments (2 races, 2 sexes, 2 age groups, 2 coronary risk levels, 3 types of chest pain, 3 types of exercise test results) are embedded in 8 videos of 8 patients, and each physician evaluated one of these 8 patients.

and they speak Dutch fluently with no accent. This implies that "foreignness" or minority status of the defendants can only be inferred from their darker skin complexation, but not from their names or accents.[6] As explained in Section III, a separate group of 89 students, who are not involved in the experiment, confirmed that the defendants are easily identified as being white or minority by the pictures of their faces.

The results reveal bias in the conviction decision against minorities, equally driven by white and minority evaluators. That is, minority defendants have a higher probability of conviction regardless of whether they face a white or minority evaluator. This finding suggests statistical discrimination against minorities by minority evaluators, under the assumption that taste-based discrimination would not operate against one's own in-group. White evaluators' higher propensity to convict minority defendants, on the other hand, may be the result of the combination of both statistical discrimination and taste-based discrimination against minorities.

In the sentencing phase we find in-group favoritism in the assigned prison sentence and fine is driven by white evaluators. Because the bulk of the evaluators in the sample is white, this translates into racial bias against minorities in prison sentences and fines. This racial bias in sentencing is a reflection of taste-based discrimination because statistical discrimination is not relevant at the sentencing stage.[7]

We also analyze whether disparate treatment of defendants is impacted by evaluators' trust in the judiciary, by their trust in police, and by their concerns about terrorism and immigration. We find limited evidence that such concerns influence racial biases in these judicial decisions, suggesting that the source of the bias may be deep-rooted.

Section II describes the general idea and the contribution of the paper. Section III provides the details of the experimental design and the data. Section IV presents the descriptive statistics,

[6] The body language was also similar between the defendants, as we discuss later in the paper.

[7] In contrast, both statistical discrimination and taste-based discrimination could be relevant in the decision to determine guilt vs. innocence.

Section V includes the econometric framework specifications and Section VI presents the results. Sections VII includes extension and robustness checks. Section VIII presents the analysis of attorneys' decisions. Section IX analyzes the question of whether the conviction and sentencing decisions of the evaluators are impacted by the extent to which the evaluator believes that immigration is a problem, terrorism is a problem, or whether the police or the judicial system can be trusted. Section X is the conclusion.

II. The General Idea and the Contribution of the Paper

Imagine we are watching a trial. The district attorney presents his case the way he sees appropriate. The defense attorney defends her client with the knowledge and the skill she possesses. The facts of the case are presented and debated by both parties. The defendant, who is sitting in front of the judge is listening to the arguments and counter-arguments made by the prosecutor and by his defense attorney. The defendant is a white man. At the end of the trial the judge makes a decision based on the arguments presented. She renders a verdict about conviction/acquittal; and if she finds the defendant guilty, she assigns a prison sentence and/or a fine.

Now imagine we go back in time and watch the same exact trial. The same prosecutor, the same defense attorney, the very same arguments, the exact same words, the exact same body language. Everything that took place in the courtroom is precisely the same to the last word and to the last body movement, with one difference: the defendant is someone else. He is not a white man, but he has dark skin. Would the decision of the judge about the verdict, the prison sentence length and the fine be different? Are these judicial decisions influenced by whether or not the judge and the defendant have similar racial/ethnic backgrounds? If so, are these influences stronger for white judges or minority judges? If there are race-driven effects on these decisions, are they impacted by whether the judge believes that immigration is a problem, terrorism is a problem, or whether the police or the judicial system can be trusted? These are the questions we try to answer in this paper.

Of course, if the judge were to watch the same trial the second time (the only difference between the two versions being the defendant's skin color), she would immediately recognize that this was the same case she adjudicated before. Therefore, while the first judge watches the trial with the white defendant, we ask another judge to watch the same trial with the defendant who has dark skin. To be exact, we have a total of 192 observers who watch six different criminal trials from the view point of the judge in the courtroom (from the bench of the judge). Each of these six trials has a different defendant who faces a different criminal charge with different circumstances. Each of these six trials has two versions. The only difference between the two versions is that while the defendant in Version 1 of each trial is a young white man, the defendant in the Version 2 of the same trial is a young man with darker skin color. We mix these trials and then randomly assign them to these 192 evaluators, who act as judges, making sure that everyone watches all six trials in random order, and that everyone watches three trials with white defendants and three other trials with dark-skin defendants. Half of the evaluators watch Version 1, while the other half watched Version 2 of each VR video.

At the conclusion of each trial, each evaluator renders a verdict on guilt vs. innocence and assigns a sentence if s/he finds the defendant guilty. We compare the verdicts and the prison sentence lengths and fines between the evaluators who watched the same exact trial but faced a different defendant. This design enables us to identify the impact of the race of the defendant on judicial decisions, *holding constant everything else that can influence the decisions.* Furthermore, because we have information on the evaluators who made judgments on the defendants, we investigate whether the evaluators hand down more or less severe verdicts and sentences if they themselves are a minority in the society.[8]

The key component of this design is that the evaluators, as judges, *observe the exact same activity in the same exact courtroom*. That is, they hear the exact same arguments made by the prosecutor and the defense attorney the way their words are spoken, and they observe the same body language and movement in the courtroom. Half of the evaluators, however, see a young white defendant in front of them, while the other half sees a young dark-skinned defendant.

[8] This component could not be done in the second group of evaluators because this group consisted of 36 attorneys and there was no minority among them.

This is achieved by VR technology. As explained in detail in Section 3, this technology allows us to video-record the courtroom activity in 3D. It also allows us to replace an individual in this VR video (the defendant) with another individual who is video-recorded separately. In the end, this production generates two identical VR videos with one difference: version one contains the white defendant, and version two contains the defendant with dark skin color.[9]

Figure 1 displays a side-by-side image of version 1 and version 2 of a particular scene from one of the trials. The version in the top frame involves a white defendant and the version in the bottom frame involves a minority defendant. In each frame, the person on the right is the prosecutor who is presenting his case. The defendant is sitting in the middle; and the person on the left, behind the defendant, is the defense attorney. This picture depicts how the judge (the evaluator) observes the trial, although watching the Virtual Reality videos using 3D headsets produces a much more realistic image of the actions and sounds of the courtroom.

Two short segments of this trial (shown in Figure 1) titled "Split screen Comparison of Version 1 and Version 2 of Video No 4" are at the link: http://proficient.ninja/splitscreen/. This particular clip presents a 40-second segment of the fourth trial shown (Video No 4), but it displays two versions of the same trial simultaneously on the split screen. The top segment of the screen displays the trial with the white defendant and the bottom segment displays the same trial with the minority defendant. Every detail in both videos is identical with the exception of the defendants. The second video at this link presents the full version of the same trial with the minority defendant (Video Number 4, Version 1), and the clip below (Video No 4, Full Version 2) displays the same trial with the white defendant.

Another short video-clip at the link http://proficient.ninja/uhasselt/ provides a glimpse of a different trial (note that the prosecutor and the defense attorney as well as the defendant are different). This video is titled "Video No 1, Version 1" to indicate the first trial (Video No 1)

[9] 3D Virtual Reality videos offer advantages compared to regular 2D videos. Once the participants put on the headsets they experience "immersion" into the actual environment, meaning that they feel part of the environment in which they are being placed, and they no longer get sensorial stimuli from their actual physical location. Also, the participants can look around in the 3D Virtual Reality videos in a 360 degrees environment instead of viewing the setting from a fixed focal point in a 2D video. Furthermore, the participants also experience "presence", meaning that they actually feel like being present in the environment because the sense of reality is increased significantly compared to 2D videos (Sadowski and Stanney 2001, Slater and Wilbur, 1997).

with the white defendant (Version 1). Two short segments are included in the clip. The first segment shows a section from prosecutor's opening remarks; and the second part shows a segment in which the defense attorney speaks. Pictures from other trials can be found at this link by scrolling down.

The prosecutors and the defense attorneys in all videos are actual prosecutors and actual defense attorneys who practice criminal law in the province of Limburg, Belgium. We used a real courtroom in Hasselt, Limburg to shoot the VR videos of the trials. The VR videos are shot from the bench of the judge so that the evaluators can observe the trial from the vantage point of the judge. There is an actual judge sitting in the bench during the trials, not seen by the evaluators who watch the VR videos. But of course the judge is seen by the defendant, by the prosecutor, and by the defense attorney during the trial. This forces all parties in the courtroom to speak towards the judge (towards the bench). Thus, when the evaluators watch the videos using 3D headsets, they feel that they are being spoken to by the prosecutor and the defense attorney. The actual judge, however, speaks very little during the trial, just to direct the proceedings. Only the defendants in the court room are actors. To represent the racial and ethnic mix of minorities in Belgium, the actors who represent minority defendants are either of Moroccan or Turkish decent.

The first group of evaluators, who watch the VR videos and make decisions as a judge, are a mixture of undergraduate and master's economics students, and masters' law students from Hasselt University in Belgium. It can be argued that the decisions of undergraduate and master's economics students and even those made by master's level law students, although informative, do not correspond to the decisions made by actual judges. It is important to underline that in Belgium a law graduate is integrated into the judicial system very quickly. Specifically, a law graduate can begin his/her career path towards becoming a judge with a judicial internship, provided that s/he has one year of experience in the field (which can consist of an internship at the Bar or any other legal function). During the judicial internship, the intern is active in at least one bench of a first instance court, where s/he has the assignment to support the judges, participate in judicial deliberations and draft verdicts. Thus, law graduates in Belgium can start

a judicial internship as early as one year after graduation, and during this internship they are expected to make judicial decisions in court cases. After three years of internship (i.e. four years after graduation) law graduates can become a full-fledged judges.

Nevertheless, to address the question of whether students' decisions are different from those who are actively involved in the judicial system, we repeated the experiment with practicing Belgian attorneys who specialize in criminal law. The details of data and the procedures are provided in the next section and in Appendix 1.

III. The Design of the Study and the Data

We are granted permission by the district attorney of Limburg, Belgium to use actual criminal cases as well as actual prosecutors and an actual courtroom. Upon discussions with two prosecutors we decided to focus on burglary and assault cases because such cases provide substantial discretion to the judge about sentencing, and no special background or additional expert information is needed to deliver a verdict (which would be the situation in cases such as medical liability).

Three actual burglary and three actual assault cases are selected from the court archives or from the docket. Although we did not change any details of the cases, we altered the names of the actual victims and witnesses to guarantee anonymity of the real cases. As explained below, synopses of actual case files are provided to the evaluators before the beginning of the experiment so that they could assess the background of each case, the forensic evidence, the background of the defendant, and so on. We did not provide the name of the defendant to evaluators in these case files to avoid any signaling about the race of the defendants. In the videos, the defendant is always referred to as 'my client' (by the lawyer) or as 'the defendant' (by the prosecutor).

For burglary cases we changed the location of the event to eliminate the possibility of an evaluator being familiar with a particular burglary incident in a particular neighborhood. Given

that the actual defendants were in some cases older than our actor-defendants, we changed the age of the defendant in the case files that are provided to the evaluators to match the age of the defendant-actors in our courtrooms.

Two prosecutors who specialize in burglary and assault cases volunteered to participate. We contacted a number of law firms that specialize in criminal cases to be part of the experiment. After interviews, we selected three lawyers (two female and one male) who have experience in both burglary and assault cases. The lawyers and the prosecutors were given the case files one week before the shooting of the videos and were asked to prepare the cases similar to an actual court trial. The lawyers did not talk to the prosecutors before the shooting to keep the trial environment as realistic as possible.

In Belgium the judge is in control of the court hearing although his/her actual role is limited. The judge first asks the defendant if he understands the charges against him/her and whether the defendant pleads guilty or not guilty. In each of our videos, the defendants indicate that they understand the charges and that they plead not guilty. The judge then allows the prosecutor to start with his statement. Once the prosecutor is finished, the defense attorney can launch his/her pleading. In an actual hearing, the judge sets the time frame for her verdict (normally four weeks after the conclusion of the hearing) and she could potentially ask the defendant some questions during the trial. Our set up (the evaluators watching the VR videos with headsets) does not allow the evaluators to interrogate the defendant. It is, however, not uncommon in Belgian criminal proceedings for the judge not to ask any questions to the defendant.

For the sake of realism in our videos, we needed an actual judge to control the flow of proceedings. Therefore, we selected a retired judge from the region of Hasselt to be in control of the cases, as he would normally do. The judge was sitting on the bench, behind the VR camera. Thus, the evaluators who watched the videos could only hear his voice in the VR video,

but could not see him. This way the setting remained realistic and the evaluators viewed the cases from the perspective of the judge.

We needed three white defendants (each would act as the defendants in two trials) and three non-white defendants (each would take part in two trials) for our six cases. We selected male actors who lived outside of the region where the evaluators reside in order to avoid the possibility of an evaluator recognizing one of the actors. We also made sure that all actors were about the same age, and we gave them instructions on wardrobe so that the clothing worn by the pairs of actors was similar. Consistent with the demographic structure of the country, we chose actors who are of Moroccan or Turkish descent to represent minority defendants. The pictures of these actors are provided in Figure 2. Their snapshots during the trials can be found by scrolling down at the link http://proficient.ninja/uhasselt/.

A professional Virtual Reality production firm was hired, and a cameraman and a technician were in charge to ensure the technical quality of the video and the audio. A static 360 camera (OZO) was used to shoot the cases. The camera was positioned before the bench of the judge and was on the same height as the judge, which allowed the participant in the experiment to observe the court room from the exact same angle as the actual judge. All videos are shot in one court room in the main court building in Hasselt, Belgium.

We involved 156 students from the faculties of Economics and Law of Hasselt University to act as judges in these trials. The Economics students were a mix of juniors and masters' students that were enrolled in the Policy Evaluation course. The Economics group consisted of 89 students. They were randomly subdivided into four groups (because we had 25 headsets to watch the VR videos) to participate in the experiment. All four groups watched the videos on the same day, and we made sure there was no interaction between any of the four groups of students. To avoid interaction between the groups, we put them in separate rooms and let them watch a movie (or be lectured by an instructor) while one of the other groups was involved in

the experiment. They were also instructed not to inform each other via mail or phone. A professor was always supervising the group of students who were not in the lab.

Law students are enrolled in Master's degree program at Hasselt University. They participated in the experiment as part of a course on Research Methods. There were 67 law students enrolled in this course, and they all took the experiment the day after the Economics students. Law students were randomly divided into three groups and again we made sure they could not interact with the other groups as the experiment was ongoing for one of the groups.

The evaluators (students) were told that they were participating in an experiment which was an exercise in the adjudication of criminal acts. We organized a random lottery (with 20 movie tickets) for students who participated in both the experiment and the follow up survey.[10]

The follow up survey, which is completed by the evaluators over the internet, is used to gather background information on the evaluators. We decided to collect this information not on the same day of the experiment but nine days after it to minimize concerns about incorrect information being provided.[11] The survey contained 40 questions, and the key questions were sprinkled throughout. For example, we asked innocuous questions about concerns regarding unemployment, importance of family, and so on, before asking whether they think terrorism is a major problem in Belgium.

We also recruited practicing attorneys from the Flanders Bar Association. By email and by phone we contacted 250 lawyers who were listed under the category of "criminal law" on the web site of the Association. Thirty-six lawyers agreed to participate. The lawyers received the same information as the student participants and the flow of the proceedings was identical.[12]

[10] The experiment took place in November 2017.

[11] If we had asked the evaluators questions about their cultural background, about their concern regarding crime and the judicial system etc. immediately after they completed the experiment, they might have been cognizant about the purpose of the survey, and may have provided untruthful responses.

[12] The lawyers took the experiment individually during a ten-day period in July 2018.

Because the overwhelming majority of the lawyers are white we could not investigate in-group bias in their decisions, but we were able to analyze the existence of racial bias in lawyers' decisions.

The evaluators watched the VR videos using VR headsets (Oculus Rift headsets) on powerful computers with high quality graphics cards. The lab hosted 25 students at one time (See Figure 3). The evaluators were granted anonymity during the experiment and in the follow-up survey (see Appendix 1 for details). For the purpose of the experiment we developed a website that would guide the participants through the experiment. The website contained instructions for the evaluators on what to do (when to put on the headsets, when to take them off, when to read a case folder, etc.). The website also communicated with the headset so that the videos would play automatically in the correct order, and provided a platform which allowed the evaluators to submit their decisions in each case (conviction/acquittal and sentences in case of conviction). As a real judge would do, the evaluators had to first read the case folder pertaining to that particular case. The case folders were color-coded to avoid confusion. That is, each of the six cases were assigned a different color folder. The case folders contained the relevant background information about the cases, including the summary of the police report, and the criminal history of the defendant (See Appendix 1 for the details of each of the six cases, the way presented to the subjects).

After reading the case folder, the evaluator could click on a button on the screen to start the video. They had 30 seconds to put on the headsets to watch the VR video. After watching the video, they took off the headsets, and clicked on a button on the screen to go to the deliberation phase. First, they were asked to decide on conviction or acquittal. If they decided to convict, they had to assign the sentence (prison sentence and/or fine). As is the case with real judicial decision, they also had to make a decision on whether and how much to suspend the prison sentence and the fine. All of these decisions were made by clicking and typing on the screen.

The experiment lasted for two hours, with a five-minute break (interactions with other participants were not allowed) after watching the first three videos and making judgments on these first three cases. After watching and rendering decisions on all six cases, participants were presented with an overview of their decisions for the all the six cases and they could alter any decisions they made (conviction, prison term and fine) on any of the six cases before they were finished with the experiment. In the analyses we use the final decisions made on each case.[13]

If the defendants in the VR videos are not recognizable as being a member of a minority or majority group in Belgium, then our design will provide null results.[14] Put differently, if the evaluators cannot infer correctly the racial background of our defendant-actors by looking at them, then the evaluators would not assign differential verdicts or sentences even if the evaluators had racial biases. To put bluntly, the question is whether the evaluators can tell the difference between a Belgian person of European heritage and another individual who is a racial/ethnic minority with dark skin (e.g. a Belgian with Moroccan or Turkish origin)? Or, would the evaluators think that the latter person is a white European, but happens to have sun tan? The pictures of the defendants are provided in Figure 2.

[13] To investigate if the evaluators took their task seriously, we conducted a number of tests such as the analysis of the decisions by the speed of the decisions made by the evaluators, by the order in which the decisions were made, and so on. These are detailed in Section VII.

[14] Ethnicity refers to common ancestry, a perception of common history and culture. Race typically refers to a group of people defined by physical characteristics. Race classifications are not established by a set of natural or biological factors but they are human constructs where skin color, eye shape, height, hair type are used as markers to define races (Omi and Winant 1994). Race and ethnicity often overlap, and groups of people may move between categories. For example, as explained by Omi and Winant (1994), for the British, Irish were considered as a different race, although they had similar physical attributes as the British. In the U.S. Irish used to be treated as similar to the Black race. Now they are considered "white" but they constitute an ethic group. U.S. Census Bureau listed five races in 1870: White, Colored (Blacks), Colored (Mulattoes), Chinese, Indian. In 1950 there were three race classifications: White, Black, Other. In 1990 the classifications became White, Black, Asian or Pacific Islander, American Indian or Alaska Native, and in 2010 they consisted of White, Black, Asian, Hawaiian or Pacific Islander, American Indian or Alaska Native. The Census Bureau contemplated using Arab as a new racial category for the 2020 census, although the idea was not implemented.

We used a different group of 89 students who were enrolled in a freshman Microeconomics course of Business Engineering degree at Hasselt University to test whether the students can correctly identify the race of the actors used in the experiment. These 89 students have not participated in the experiment and presumably have no knowledge of the experiment. The students were in an auditorium and a large screen in the auditorium displayed defendants' pictures one at a time. Under each picture there were options to choose from regarding the racial/ethnic heritage of the person in the picture. The options were: Western European descent, Middle Eastern or North African descent, and Asian descent. Students were told that examples of Western European descent would be countries such as Belgium, Netherlands, Germany and France; Middle Eastern and North African descent would include such countries as Turkey, Morocco, Syria and Iraq, and examples of Asian descent would include countries such as China, Japan and South Korea. Each student had a hand-held device to enter their choices within 30 seconds after each picture was shown. The students were not allowed to talk to each other during this process. In addition to the six defendants in our trials, we added the picture of a young Chinese man for the students to evaluate.

The results showed almost unanimous consensus regarding the racial/ethnic heritage of the individuals in the pictures; and their heritage was identified correctly. Specifically, 97 percent of the students correctly guessed that the first minority defendant was of Middle Eastern/North African descent. All students who participated in this exercise guessed correctly the heritage of the second and third minority defendants. The students guessed with 98 percent accuracy that the first and the second White defendants were of Western European origin (Belgian), and they guessed with 100 percent accuracy that the third white defendant-actor had in fact Western European descent. Also, all students correctly identified the Chinese person as someone of Asian descent. Overall the minority defendant-actors have been identified as being a minority

with almost 99 percent accuracy, and white defendants are identified correctly as being of Western European descent with 98.5 percent accuracy.

Because we have six different cases (three burglaries and three assaults) and two versions of each case (one with a white defendant and one with a non-white defendant) we have a total of 12 videos in the experiment, and each defendant-actor participated in two videos. We made six sets with different sequences of the videos and three different combinations to make sure that each evaluator would see each defendant only once. The sets are listed in Table 1. BUR refers to burglary cases and AS stands for assault cases. The numbers in the upper right superscript refers to the case number (e.g. BUR^{1} means the first burglary case, AS^{3} represents the third assault case, and so on). The lower right-hand subscript refers to the prosecutor who handled the case. There are two prosecutors in the experiment. B refers to Bruno who handled the burglary cases, and P stands for Pieter who handled the assault cases). The lower left subscripts on the entries in Table 1 refer to the defense lawyers (SO, TO or M). The upper left superscript refers to the defendants. WH1, WH2 and WH3 stand for three white defendants, and M1, M2 and M3 stand for three minority defendants. The pictures of the defendants are provided in Figure 2.

Each evaluator watched one of the six sets depicted in Table 1. Thus, each one of the three burglary cases (BUR^{1}, BUR^{2}, BUR^{3}) as well as each of the three assault cases (AS^{1}, AS^{2}, AS^{3}) is watched by each evaluator. There are three different sequences with which the cases are watched. In sequence 1, the first three burglary cases are presented, and they are followed by the three assault cases. Sequence 2 presents these cases in reverse order: AS3, AS2, AS2, followed by BUR3, BUR2 and BUR1. Sequence 3 presents a burglary case, followed by an assault case, followed by another burglary case, and so on. Each sequence has two versions. If a case includes a white defendant in version 1, the same case includes a minority defendant in version 2. For example, consider Set 1 and Set 2. In Table 1 the videos are presented in the

same sequence (Sequence 1) in both Set 1 and Set 2. The only difference between Set 1 and Set 2 is that, if a particular case involves a white (minority) defendant in Set 1, the same case involves a minority (white) defendant in Set 2. Similarly, Sets 3 and 4 are identical with the exception of the race of the defendants. The same is true for Sets 5 and 6. To be more specific, consider the first video in Set 5: ${}^{\mathrm{WH2}}_{\mathrm{TO}}\mathrm{BUR}^{3}_{\mathrm{B}}$, which represents the third burglary case, prosecuted by B, where the defendant is WH2, who is defended by lawyer TO. The first video of Set 6 is ${}^{\mathrm{M2}}_{\mathrm{TO}}\mathrm{BUR}^{3}_{\mathrm{B}}$, which signifies the same exact case with the same prosecutor and the same lawyer, but with the minority defendant M2.

IV. Descriptive Statistics

Table 2 presents the descriptive statistics related to student evaluator attributes. *Minority Evaluator* takes the value of one if either the mother or the father of the evaluator was born outside of Belgium, and if that country is outside of the European Union (EU).[15] Non-EU countries that are revealed by evaluators as mother's country of origin are India, Morocco, Turkey, South Korea, Zimbabwe and Armenia. Non-EU countries that are revealed as fathers' country of birth were Morocco, Turkey, Zimbabwe, Armenia and Japan.[16] Alternatively, we determined if the evaluator was a minority in Belgium based on the answers to the following question. "Do you or anyone in your household speak another language other than Dutch?"[17] In this alternative definition, we coded the evaluator as a minority if he/she indicated that either Arabic, Armenian, Turkish or Shona is spoken at home.[18]

[15] If the parent was born in a country which is part of the EU, such as Germany or France, the evaluator is not considered as a minority in Belgium.

[16] Variation in this definition did not alter the results. For example, considering those evaluators whose ancestry is Japanese or Korean as somebody White, rather than minority (Asian) had no impact on the results. Similarly, narrowing the definition of *Minority Evaluator* to those whose mother or father is from Morocco, Turkey, Armenia or India did not influence the results.

[17] Hasselt University is located in Flanders, the Dutch-speaking part of Belgium.

[18] We did not consider evaluators as a minority in Belgium when for example French, Polish or Italian is spoken at home.

Using the same survey that was registered nine days after the experiment, we also measured evaluators' beliefs and attitudes on a number of social issues. The dummy variable *Trust in the Legal System* takes the value of 1 if the individual assigned a value of 7 or higher to the question "On a score of 0-10, how much would you say you personally trust the legal system? 0 means you do not trust the legal system at all, and 10 means you have complete trust." A similar question is used to gauge trust in police. "On a score of 0-10, how much would you say you personally trust the police? 0 means you do not trust the police at all, and 10 means you have complete trust." *Trust in Police* takes the value of 1 if the person gave a rating of 7 or higher to this question.

Using a scale from 0 to 10, the evaluators could also reveal attitudes towards immigration in response to the question "Is Belgium made a worse or better place to live by people coming to live here from other countries?" Those who chose 1-6 are considered to be those who believe that Belgium is worse off because of migration (70 percent of the evaluators).[19]

The evaluators were also given the question "To what extent do you think it is important that Belgian society pays attention to terrorism?" and were asked to reveal their beliefs about this statement as (1=very unimportant, 7=very important). If the person chose 6 or 7, we identified him/her as agreeing that terrorism is a major problem in Belgium.

Table 3 presents the descriptive statistics of the decisions made by evaluators. Row (1) displays information about the conviction decision by the race of the defendant [Columns (1) and (2)], as well as by race-matching between the defendant and the evaluator [Columns (3) to

[19] The three questions on the trust in the legal system, trust in the police and the extent to which immigration is a problem in Belgium were asked in the same way and using the same scale in the European Social Survey (ESS). Using the ESS survey years of 2016 we found that in the overall population of Belgium 36 percent trust the legal system, 58 percent trust the police, and 75 percent believe that Belgium is made a worse place to live because of immigration. The rates of trust in the legal system and trust in the police are similar between the overall Belgium population (ESS survey data) and those observed among our evaluators, reported in Table 2. The percentage of those who believe that immigration made Belgium a worse place to live is lower among our student evaluators in comparison to the overall population (58 percent vs. 75 percent).

(6)]. For example, columns (1) and (2) reveal that the conviction rate among minority defendants is 11 percentage points higher in comparison to the conviction rate of white defendants (0.79 vs. 0.68) and that the difference is statistically different from zero. Columns (3) and (4) show that minority defendants are more likely to get convicted if they face a white evaluator. Columns (5) and (6) indicate that the same is true for minority defendants if they faced a minority evaluator. That is, minority defendants are 17 percentage points *more* likely to get convicted in comparison to white defendants if both groups of defendants are judged by a *minority evaluator*.

Row (2) of Table 3 displays the average prison sentence imposed on defendants if they were found guilty. Part of the prisons sentence can be suspended at the discretion of the evaluator. Row (3) presents the average suspended sentence among various groups. The difference between the initial prison sentence and the suspended sentence is the actual, effective prison term faced by the convicted defendants, displayed in row (4). For example, column (1) shows that convicted minority defendants receive an average of 10.2 months prison time (row 2), and that 5.8 months of this initial sentence is suspended (row 3). Thus, row (4) and column (1) show that the effective prison term among minority defendants is 4.4 months, on average.

Columns (3) and (4) and row (4) of Table 3 show that minority defendants receive effective prison sentences that are one month longer than those imposed on their white counterparts if the defendants faced a white evaluator (4.6 months vs. 3.6 months). This same is true in reverse for minority evaluators. Row (4) and columns (5) and (6) show that minority evaluators assign substantially longer sentences to white defendants in comparison to minority defendants. This suggests that evaluators provide preferential treatment in prison sentencing of defendants who are of the same race, which reflects in-group bias in effective prison term.

Notice that in-group bias also exists in the initial prison sentencing (Row 2). Minority evaluators hand down sentences to white defendants that are about 3 months longer (Row 2,

columns 5 and 6). White evaluators assign one-month shorter initial sentences to whites (9.5 vs. 10.5 months) although this difference is statistically different form zero only at the 11 percent level. Row (3) shows that there is no in-group bias in suspended sentences. This indicates that the in-group bias generated in initial sentencing is carried over to the final, effective prison term, displayed in row (4).

An interesting aspect of row (4) is that average effective prison terms is not significantly different between white and minority defendants (columns 1 and 2). The average effective prison term is 4.4 months for minorities and it is 3.9 months for whites but there is no statistically significant difference between the two averages.

Rows (5) to (7) of Table 3 reveal in-group bias in fines. White evaluators assign lower fines to white defendants, and minority evaluators assign lower fines to minority defendants. Furthermore, both types of evaluators suspend a larger chunk of that initial fine for defendants of their own race. This translates into effective fines that are lower for defendants who face an evaluator of their own race, although the differences are not statistically different from zero.

The statistical insignificance in the difference in sentencing between white and minority defendants (Table 3: columns 1 and 2, rows 4 and 7) is misleading because racial bias in the conviction decisions generates a selected sample of defendants who are sentenced. Note in Table 3 that while 370 minority defendants are convicted and sentenced, there are only 317 white defendants who are sentenced. These 53 "excess" minority defendants end up in the sentencing phase because of the 11 percentage point difference in the conviction rates between the races (0.79 vs. 0.68).[20] We present results both ways: i) that account for this selection bias and ii) that ignore the selection bias.

[20] The same point is made by Depew et al. (2017). Because the sequential decision of conviction and sentencing of the defendants by the same judges generates sample selection in the sentencing phase, the authors analyzed the sentencing outcomes of those defendants who pled guilty (as opposed to those who are convicted and then sentenced by the same judge).

V. Econometric Analysis

In-group bias

To test for the existence of in-group bias we make use of information obtained from the evaluators regarding their cultural heritage. This information allows us to estimate Equation (1A) below.

$$(1A) \quad Y_{cj} = \alpha_1 + \beta_1 \text{Minority Defendant}_c + \gamma_1 \text{ Minority Evaluator}_j + \delta_1 \text{ Minority Defendant}_c * \text{Minority Evaluator}_j + \mathbf{X}_j\Theta + \mu_c + \omega_{cj}$$

where Y_{cj} stands for the outcomes related to case *c*, judged by the evaluator *j*. The first outcome is an indicator to represent whether the defendant is convicted by evaluator (*j*). The second and third outcomes are the prison term and the amount of monetary fine imposed on the defendant if he is found guilty. *Minority Defendant* is a dummy variable which takes the value of one if the defendant in case *c* (in the VR video of case *c*) is a racial Minority. Recall that a particular case is identical across all evaluators who watch the VR video of that case, with the exception of the skin color of the defendant. In other words, all evaluators watch the same video of a particular trial (indexed by *c*), except that some evaluators are randomly chosen to watch the version with white defendant, while other subjects are randomly assigned to the version that includes the minority defendant.

In Equation (1A) Minority Evaluator takes the value of 1 if the evaluator, who makes the conviction and sentencing decisions, is a minority him/herself in Belgium. As explained in the data section, we measure this variable in two ways: i) whether at least one of the parents of the evaluator was born in a developing country in the Middle East or North Africa, and ii) whether a language such as Arabic, Turkish or Armenian is spoken in the household of the evaluator.

Six courtroom hearings (c=1, 2,…6) are watched and judged by each evaluator (j). Three of these cases involve burglary, and the other three are assault cases. In some specifications we include case fixed effects (μ_c), in others we control for the type of crime on which the trial was

based (assault vs. burglary). Vector X includes attributes of the evaluators such as their sex and whether they are law students or economics students. Although all evaluators watch the same cases, the order in which these cases are watched is different by design. For example, some evaluators watched Case Number 1 (the hearing of the first burglary case) as their first video, while other evaluators watched the same case as their sixth and the last video. As Table 1 demonstrates, evaluators who were randomly assigned to Set 1 or Set 2 watched these six cases in a particular sequence (Sequence 1), while those who were randomly assigned to Sets 3 or 4 watched the same videos in a different order (Sequence 2); and Sequence 3 is the third sequence in which the videos are watched by evaluators. Vector X contains Sequence1 and Sequence2 which are two dummy variables that control for the sequence in which the videos are watched.

In Equation (1A) γ_1 captures the difference in outcomes between white defendant-minority evaluator pairs and white defendant-white evaluator pairs, and $\gamma_1+\delta_1$ captures the differential impact on sanctions assigned on minority defendants by minority versus white evaluators. Thus, δ_1 in Equation (1A) is the difference-in-difference estimate, which signifies the differential decisions of minority evaluators versus white evaluators in their treatment of minority defendants over white defendants. Put differently, if δ_1 is different from zero, this is evidence for in-group bias.

Inclusion of evaluator fixed effects to specification (1A) could be important in the analysis of data from actual trials because judges who preside over those hearings may embody unobserved (to the analysts) attributes that may confound the results. For example, as explained in Depew, Eren and Mocan (2017) if judges are assigned to jurisdictions non-randomly (e.g. by an election, voted by local residents of the jurisdiction) and if judge race is correlated with judge leniency in sentencing, judge fixed effects are important to control for. This is not a concern in our paper because, by design, there is no correlation between evaluator race and the attributes of the cases. Similarly, defendant's race is orthogonal to case and evaluator characteristics by

design. Nevertheless, we also estimate (1A) by adding evaluator fixed effects, which produces Equation (1B).

(1B) $Y_{cj} = \alpha_2 + \beta_2$ Minority Defendant$_{cj}$ + δ_2 Minority Defendant*Minority Evaluator

+ Cc Ω + ηj +υc+ ξcj

Evaluator fixed effects (η_j) in Equation (1B) implies that the Minority Evaluator dummy cannot be included. As a result, the coefficient δ_2 in Equation (1B) represents the in-group bias in terms of the impact of same-race matching between the defendant and the evaluator. For example, if Y_{cj} stands for the length of the prison sentence and if $\delta_2<0$, this indicates that the defendant receives a lower sentence if the evaluator who makes the sentencing decision is of the same race as the defendant.

If in-group bias exists, it is important to investigate whether this bias is driven by white or minority evaluators. While specifications (1A) and (1B) are important in the investigation of in-group bias, they cannot determine the source of the bias. For example, assume that the difference-in-difference coefficient δ_2 in Equation (1B) is negative. This indicates that evaluators are more lenient towards defendants of their own race, but this does not specify whether this impact is driven by white or non-white evaluators. Or, assume that δ_1 is negative in Equation (1A). This finding could emerge when i) white evaluators treat both types of defendants with equal severity, but minority defendants treat minority defendants more leniently, ii) white evaluators treat minority defendants more harshly than white defendants, but minority evaluators treat minority defendants (much) more leniently in comparison to white defendants, iii) both minority and white evaluators treat minority defendants more leniently, and so on.

Our framework enables us to decompose the in-group bias effect δ_2 in Equation (1B) to determine the source of the bias. Because the race of the defendant is, by design, uncorrelated with all courtroom attributes and with all case characteristics, and because it is also uncorrelated

with the race of the evaluator, we can run the regressions below, conditional on defendant race, to investigate the source of the in-group bias.

(2A) $Y_{cj} = \alpha_3 + \gamma_3$ Minority Evaluator $+\mathbf{X}_j\Theta + \mathbf{C}_c \boldsymbol{\Lambda} + \mu_c + u_{cj}$ in the sample of Minority defendants

(2B) $Y_{cj} = \alpha_4 + \gamma_4$ White Evaluator $+\mathbf{X}_j\Pi + \mathbf{C}_c \boldsymbol{\Xi} + \theta_c + \tau_{cj}$ in the sample of White defendants.

The in-group bias coefficient of δ_2 in Equation (1B) is equal to $(\gamma_3 + \gamma_4)$ in Equations (2A) and (2B). For example, if Y stands for the probability of conviction and if $\delta_2<0$ in Equation (1B), this implies positive in group bias. That is, evaluators are less likely to convict the defendants who are of the same race. Equations (2A) and (2B), however, provide further information. For example, if $\gamma_3<0$ and $\gamma_4=0$, this implies that the in-group bias identified in the data (by Equation 1B) is driven by the behavior of minority evaluators. Alternatively, if $\gamma_4<0$ and $\gamma_3=0$, this indicates that the in-group is due to the behavior of white evaluators. It could also be the case that $\gamma_3<0$ and $\gamma_4<0$, indicating that both groups are biased in favor of their own race.

Racial Bias

To investigate the existence of *racial bias* in conviction and sentencing decisions we run regressions of the following type:

(3A) $Y_{cj} = \alpha_5 + \beta_5$ Minority Defendant$_{cj} + \mathbf{X}_j\Phi + v_c + \varepsilon_{cj}$

Note that in Equation (1A), which analyzes in-group bias, $(\alpha_1 + \beta_1)$ represents the impact of white judges on minority defendants, and α_1 stands for the impact of a white judges on white defendants. Thus, β_1 represents the differential treatment of minority defendants vis-à-vis white defendants by white judges. Similarly, $(\beta_1 + \delta_1)$ stands for the differential treatment of minority defendants by minority judges. If *p* stands for the proportion of white judges in the sample, then

the overall differential treatment of minority defendants by all judges is equal to $[\beta_1+(1-p)*\delta_1]$, which should be equivalent to the estimated coefficient of *Minority Defendant* β_5 in Equation (3A).

Equation (3B) below is a different version of (3A) as it includes evaluator fixed effect (λ_j).

(3B) $$Y_{cj} = \alpha_6 + \beta_6 \text{Minority Defendant}_{cj} + C_c \boldsymbol{\Psi} + \lambda_j + \upsilon_c + u_{cj}$$

There are 156 evaluators in the first group (law students and economics students), and there are 36 evaluators in the second group (attorneys). In specification (3B) identification of the coefficient of the Minority Defendant dummy (β_6) is obtained from the variation of defendant race across the cases evaluated by each evaluator. The term υ_c represents case fixed-effects, and the vector **C** is a subset of vector **X** because it cannot include evaluator attributes as they are absorbed by evaluator fixed effects (λ_j).

VI. Results

In-group Bias in Conviction Decisions

Table 4A presents the estimation results of model (1A). These specifications investigate the existence of in-group bias in conviction decisions, which is identified by the coefficient of the interaction term (*Minority Defendant x Minority Evaluator*). There are 156 evaluators in the group of law students and economics students, who made decisions on conviction vs. acquittal on each of the six cases they watched in virtual reality videos, generating a sample of 936 decisions. The analyses of lawyer decisions are reported later in the paper. In column (1) of Table 4A we control for whether the case was an assault or burglary, as well as whether the decision on the case was made during the first half of the experiment or in the second half (*Early Trial*). Column (2) presents the results from another specification which includes trial fixed-effects. Column (3) adds the sex of the evaluators and whether the decision was made during

the first half of the experiment. Column (4) controls for the Sequence type (the order in which videos are watched), in addition to trial fixed effects.

No matter what specification is estimated, the point estimate of the coefficient of Minority Defendant is always around 0.11 and statistically significant. On the other hand, the interaction term (Minority Defendant x Minority Evaluator) is never significantly different from zero, indicating no in-group bias in the conviction decision. These two coefficients indicate that minority defendants have a higher probability of conviction regardless of whether they face a minority evaluator or white evaluator.

Table 4B presents the results of model (1B), which controls for evaluator fixed effects. The results are consistent with those reported in Table 4A. Regardless of how the model is estimated, the point estimate of the Minority Defendant x Minority Evaluator dummy is not statistically different from zero, indicating lack of in-group bias.

This inference is supported by the results presented in Table 4C, which displays the results based on Equations (2A) and (2B). Recall that these models help identify the source of the in-group bias (i.e. whether in-group bias is driven by white or minority evaluators). As described above, there is no in-group bias in conviction decisions, and Table 4C supports this finding. The first column of Table 4C presents the results of the regression that uses the sample of white defendants, and column (2) is based on the sample of minority defendants. As shown in column (1) of Table 4C, the coefficient of the variable White Evaluator is not different from zero, indicating that white defendants are not treated more leniently by white evaluators. Similarly, column (2) of Table 4C shows that minority defendants are not treated more leniently by minority evaluators. Thus, Tables 4A, 4B and 4C reveal that minorities are discriminated against in conviction decisions and that both minority and white evaluators are responsible for this outcome. This implies that (and is shown in empirical analysis) there is no in-group bias in convictions.

Racial Bias in Conviction Decisions

Tables 4D and 4E present the results of estimating models 3A and 3B to investigate racial bias in conviction decisions, where the outcome is the probability of conviction and the key explanatory variable is the race of the defendant. The results show that minority defendants are about 11 percentage points more likely to get convicted. This result is consistent with those reported by the in-group bias regressions (Tables 4A and 4B).[21]

These results indicate that two defendants, whose cases are identical in every respect except for their race, receive different resolutions. Specifically, the defendants face the same exact criminal charge with the same exact mitigating and aggravating circumstances. They are charged by the same prosecutor and are defended by the same attorney in the same exact way (same spoken language and same body language in the courtroom). Nevertheless, minority defendants are 11 percentage points (about 15 percent) more likely to get convicted in comparison to the white defendant, *all else being the same*. Furthermore, this differential outcome emerges regardless of whether the decision-maker is a minority or whether he/she is a member of the white majority in the population.

Recall that the racial differences in convictions cannot be attributed to the behavior of the prosecutors or the defense attorneys because their behaviors in the courtroom are being held constant. Thus, the results can be due to statistical discrimination against minorities and/or taste-based discrimination against minorities. Assuming that a group of individuals cannot have systematic taste-based discrimination against their own group (i.e. having a distaste against own group), the results suggest that minority evaluators' higher propensity to convict minority defendants is likely driven by statistical discrimination against minorities. A more nuanced

[21] In other words, because minority evaluators also discriminate against minority defendants in the conviction stage, there is no in-group bias in convictions (See Table 4A), which translates into overall racial bias against minorities in conviction (Table 4D).

explanation could be that if a minority evaluator considers a minority defendant as a representative of the entire group of minorities, the evaluator might be harsher towards that defendant for badly representing the group for just being put in front of the judge, and this sentiment can translate into convicting the minority defendant even if the case against him in the court hearing may not have been very strong.[22]

White evaluators' higher propensity to convict minority defendants, in comparison to white defendants, may be the result of the combination of both statistical discrimination and taste-based discrimination against minorities.

Sentencing Decisions: Prison Term and Fine

Those defendants who are found guilty are sentenced to a prison term and/or fine. Because there is racial bias in the conviction decision, this bias contaminates the sample of defendants who are sentenced. More specifically, the existence of racial bias in convictions (Table 4D) implies that some minority defendants, who should not have been convicted, are nevertheless found guilty because of racial bias. If these individuals represent borderline cases in the conviction-acquittal decision, they may receive lenient punishment in the sentencing phase. In this case, OLS estimates of the racial bias in sentence length and fine would be biased downwards. To account for such selection, we follow Hoffman and Oreopoulos (2009) and Angrist et al. (2006) and trim the sample of those who are found guilty.

[22] In a different context, Corell et al. (2002) and Correll et al. (2007) used subjects such as police officers and college students in videogame-like simulations, where the subjects viewed a series of images of black or white individuals, sometimes holding guns or other objects. Subjects' decisions to shoot or not to shoot these targets revealed racial biases in the speed with which the decision are made, and both white and black subjects displayed the same shooter bias against black targets. Subjects were faster when shooting an armed black man in comparison to an armed white man; they were more careful in not shooting an unarmed white man in comparison to an unarmed black man. List and Price (2009) reported that minority donors are less likely to make a contribution for a charity during a door-to-door fundraising drive if the solicitor is also a minority, and that the size of the gift is lower, conditional on the decision to give, suggesting lack of trust.

In Table 5 we display the conviction rates of white and minority defendants by the race of the evaluator they faced. White evaluators made 411 decisions on minority defendants and they made another 411 decision on white defendants. These white evaluators convicted minority defendants 78 percent of the time, while they convicted white defendants with 68 percent probability. This generated 322 convicted and sentenced minority defendants, but only 279 white defendants who are found guilty and then sentenced. Put differently, the 10 percentage point difference in the conviction rates between white and minority defendants (0.78 vs. 0.68), which is due to racial bias of white evaluators, generated 43 excess minority defendants to be sentenced by white evaluators (322-279=43).

As shown in Tables 4A-4E, and as Table 5 also demonstrates, minority evaluators too are biased against minority defendants in their conviction decisions (conviction rates of 0.84 vs. 0.67 as shown in columns (3) and (4) of Table 5). This has created 10 excess minority defendants (48-38=10) sentenced by minority judges. The trimming procedure allows us to determine these marginal defendants and drop them from the sample. This is achieved by analyzing the sentence distribution of all 322 minority defendants who are convicted by white evaluators and by dropping 43 of these who received the lowest sentences from these white evaluators. Among the 322 minority defendants who are found guilty by white evaluators, there are exactly 43 defendants for whom both the effective prison sentence is zero and the effective fine is zero. That is, white judges assigned prison terms and fines, and then suspended all of the prison term and the entire fine in these 43 cases. We applied the same procedure to drop the 10 minority defendants (to bring down the number of sentenced minority defendants to the number of sentenced white defendants: from 48 to 37) who were convicted and then sentenced by minority evaluators: those who are at the far left tail of the sentence distribution.[23] We report

[23] Six minority defendants who are convicted by minority evaluators had zero effective prison time and zero effective fine assigned to them. These are the minority defendants with no effective sentences, and they are dropped. Among the remaining convicted minority defendants, we dropped those who had zero effective prison time coupled with lowest effective fines (which were 100 and 300 Euros).

the results based both on the trimmed sample as well as those based on the untrimmed sample (the entire group of convicted defendants)

Table 6A presents the estimation results of the in-group bias analysis (Equation 1A) where the outcome is the effective prison sentence. Recall that effective prison sentence is the actual prison sentence imposed on the defendant, which is the difference between the initial sentence handed down by the evaluator and the suspended sentence (see Table 3). This sample includes those defendants who are found guilty. Panel A of Table 6A presents the results that employ the trimmed sample and Panel B displays the results that are based on the whole sample (the entire group of convicted defendants). The results reveal the existence of in-group bias in prison sentencing. The estimated coefficient of the interaction term *Minority defendant * Minority Evaluator* is negative and significantly different from zero in every specification in Table 6A.

Note that because there is *no in-group bias in conviction decision* (see Tables 4A and 4B), the results pertaining to in-group bias in sentence length are very similar between panels A and B of Table 6A. In other words, *the lack of in-group bias in convictions* implies that any in-group bias effect in the second-stage decisions (prison time and fine) can be unbiasedly estimated using either the trimmed or the untrimmed sample.

The magnitude of the coefficients in Table 6A indicate that if the defendants are sentenced by an evaluator of their own race, they receive prison sentences that are almost 4 months shorter, on average. This is a big impact as the average sentence length is 4 months. Regressions that included evaluator fixed effects provided the same conclusion (Appendix 2, Table A2-1).

Table 6B presents the results of the analyses that investigate the source of the in-group bias. Here, effective sentence lengths are regressed on the race of the evaluator, conditional on the race of the defendant (see Equations 2A and 2B). Columns (1) and (2) employ the trimmed sample and column (3) and (4) use the entire (untrimmed) sample of convicted defendants. Note that Column (1) of Table 6B uses the trimmed sample of white defendants, and column (3)

employs the untrimmed sample of white defendants. The sample sizes are the same between these two samples and therefore the results in columns (1) and (3) are identical. This is due to the fact that in case of white defendants, trimmed and untrimmed samples are one and the same, because the sample is trimmed by eliminating "excess" minority defendants who were convicted because of their race by both white and minority evaluators. Columns (2) and (4) of Table 6B show that in the sample of minority defendants, if the evaluator is also a minority, the prison term is 1 to 1.2 months shorter, although the impact is not statistically different from zero. Columns (1) and (3) display the results of the same analysis in the sample of white defendants who were found guilty. Convicted white defendants receive sentences that are about 2.8 months shorter if they face an evaluator who is also white. These results suggest that the in-group bias in the assignment of prison sentence is primarily driven by white evaluators, although a test for the equality of the two effects cannot be rejected at conventional levels of significance.

Table 6C presents the estimation results of the *racial bias* in the assigned effective prison term. Trimming the sample of convicted defendants is expected to eliminate the marginal minority defendants who should not have been found guilty in the absence of racial bias in convictions. Thus, it should eliminate the downward bias in the estimated effect of the defendant's race on prison sentence. As expected, the coefficient of Minority Defendant is larger when the regressions use the trimmed sample in Panel A of Table 6C. The results indicate that minority defendants receive about 0.7 months longer prison terms when the racial selection bias in conviction is not adjusted for (panel B of Table 6C), but that minority defendants receive 1.3 months longer sentences (about 30% from the mean) if selection is accounted for.[24] Models with evaluator fixed effects provided the same inference (Appendix 2, Table A2-2).

[24] As discussed earlier, the coefficient of interest *(Minority Defendant)* in Table 6C is also recoverable from the two coefficients obtained from the in-group bias regressions of Table 6A: *(Minority Defendant)* and *(Minority Defendant* x *Minority Evaluator).* For example, using column 4 of the trimmed sample of Table 6A, 1.750-(3.783)*0.12 is equal to 1.296 (where 0.12 is the proportion of minority evaluators in

In the assault and burglary cases the Belgian criminal code provides the option for the judge to assign a fine in addition to a prison sentence, or in certain circumstances just a fine without a prison sentence. Thus, the evaluators can assign a fine to the convicted defendants in addition to a prison term if the evaluators find it appropriate. Table 7A presents the results of in-group bias analyses pertaining to imposing a fine. The dependent variable is the logarithm of the fine imposed on the defendant.[25] The coefficient of the interaction term (*Minority Defendant x Minority Evaluator*) is negative in all specification and the point estimate is around -1.0, indicating that defendants who have the same race as the evaluator receive fines that are 69 percent lower.[26]

Models reported in Table 7B reveal that this in-group bias effect is driven by white evaluators. Convicted white defendants receive fines that are 78 percent lower if they are evaluated by a white person (the coefficient is *White Evaluator* is -1.387), but the impact of the evaluator's race on fine is not significantly different from zero in the sample of minority defendants. Models with evaluator fixed effects can be found in Appendix 2, Table A2-3.

Table 7C displays the regression results that investigate the existence of *racial bias* in fines. The results in Panel B, that are based on the entire sample, indicate that minority defendants receive fines that are 40% higher than white defendants who are identical in all respects other than race. Panel A, which reports the results based on the trimmed sample, reveal that correcting the sample selection (stemming from the bias in conviction decision), increases the coefficient of interest to 1.023, which implies that minority defendants receive fines that are 172 percent higher. Models with evaluator fixed effects can be found in Appendix 2, Table A2-4.

---

the sample), which is the same as the coefficient of Minority Defendant in column (4) of the trimmed sample in Table 6C.

[25] Because there are some zeros in effective fines, we added one Euro to effective fines assigned by the evaluators.

[26] The impact is calculated as $exp\{\beta - 0.5 Variance(\beta)\} - 1$, where β is the estimated coefficient, and $Var(\beta)$ is its variance (Kennedy 1981).

In summary, the results of Tables 6A-7C reveal in-group bias in prison sentence and fine that are assigned to convicted defendants, and that this in-group bias is driven by white evaluators. Because the majority of the evaluators in the sample are white[27], the in-group bias in sentencing translates into overall racial bias in sentencing.[28]

Any bias that arises in the sentencing stage would be a reflection of taste-based discrimination. This is because the defendants who are sentenced have been found guilty already, and any statistical discrimination that may have impacted the conviction decision is not relevant at the sentencing stage. This implies that the racial bias against minority defendants at the sentencing phase is the result of taste-based discrimination by white evaluators.

VII. Extensions and Robustness

We used an alternative measure to determine the cultural background of the evaluators. Instead of making use of information on parents' country of origin, we determined whether the evaluator is a minority in Belgium by using information on the language used in the household (see the Data Section on details). Using this alternative indicator of minority status provided very similar point estimates.

The unconditional mean of the prison sentence and the fine assigned by the evaluators are smaller than their respective variances and the distributions are left-skewed. Therefore, we estimated the prison term and fine regressions using negative binomial models, which provided the same inference. For example, Table 6C shows that minority defendants receive prison sentences that are 1.3 months longer, which is about a 32 percent increase from the mean sentence length. Negative binomial models provided an estimate of this racial bias of 38 percent. Similarly, Panel A of Table 7C shows that minorities receive fines that are 172 percent higher

[27] Eighty-eight percent of the evaluators are white (see Table 2).

[28] Recall that, as described at the end of section V, the racial bias coefficient β_5 (the coefficient of Minority Defendant in Equation (3A) is equal to $\beta_1 + p\delta_1$, where β_1 and δ_1 are the relevant coefficients from the in-group bias regression (1A), and p is the proportion of white evaluators.

than their white counterparts (implied by the coefficient of 1.023). The negative binomial model reveals a smaller, 30 percent effect of the race of the defendant on the amount of the fine assessed.

During the experiment the evaluators had the option (as the actual judicial procedure allows for) to reflect and to revise their original decisions on conviction, prison sentence and fine. There are 156 evaluators who made 2,310 decisions (936 decisions on whether to convict, and 687 decisions on prison sentence and fine on those who are found guilty). We divided the sample into two groups: those evaluators who never altered their first decisions (90 evaluators) and those who made at least one change in their decisions (66 evaluators). Changing a decision may indicate that the evaluator contemplated more carefully about the case and therefore felt the need to revise his/her original decision. Alternatively, if an evaluator was very deliberate in watching the presentations of the prosecutor and the defense attorney and if the evaluator read the case file carefully, he/she did not have the need to go back and revise the original decisions about conviction and/or sentencing. It is also possible that chancing a decision at the end of the experiment may indicate that the evaluator has realized the purpose of the experiment, and as a result he/she went back to revise at least one of the decisions he/she made during the experiment. In this case, any statistically significant effect of racial bias would disappear or would be smaller in the sample of evaluators who changed at least one of their decisions. In any case, it is interesting to investigate whether the results are different between these two groups of evaluators.

To investigate whether the evaluators took their task seriously, we analyzed the time they spent in making their decisions. Figure 4 displays the distribution of total time (total time spent for 6 trials) spent by evaluators on decision-making. For example, 500 seconds means that a case took on average 1 minute and 23 seconds to decide for the evaluator after he/she completed

watching the VR video of the case.[29] Recall that the evaluators read the case files prior to watching the VR videos of the case. The case file contains information about the case (police report, etc., as well as the sentencing guidelines for that particular crime. See Appendix 1). A quick decision after reading the case files and after watching the trial may imply that the evaluator watched the trial carefully and formed an opinion during the trial, and did not have to think long about the verdict and punishment. Alternatively, a quick decision may indicate that the evaluator did not pay attention to the case and made a quick and haphazard decision. To investigate the sensitivity of the results to decision time, we dropped from the sample the 25 percent of fastest evaluators and the 25 percent of the slowest evaluators and re-estimated the models. We repeated the exercise by dropping the slowest and fastest 30 percent, 20 percent, and 15 percent of the distribution of cases, and obtained very similar results.

Similarly, the evaluators may have gotten fatigued during the experiment and they may have lost their concentration towards the end. If this is the case, the decisions made later during the experiment should be less careful and more noisy. Alternatively, they may have realized the purpose of the experiment, and may have made their decisions accordingly. To investigate this point, we analyzed the first three decisions and the last three decisions made by the evaluators to analyze any differences in the results between the groups. Finally, we re-estimated the models using the sample of male or female evaluators and using only law students or only economics students.

Table 8 summarizes the results obtained from these analyses. *Overall Bias* indicates the estimated coefficient of *Minority Defendant* in the relevant regression. In-group bias pertains to the coefficient of *Minority Defendant* x *Minority Evaluator*. Column (1) displays the estimates obtained from the entire sample as presented in Tables 4A-4E. Columns (2) and (3) report the results based on the first three and last three decisions, respectively, of each evaluator.

[29] This is total time spent to make the decision on guilt/innocence, and on prison sentence and fine.

Column (4) presents the estimates related to cases in which the evaluators has modified at least one of his/her initial decisions. Column (5) pertains to the sample of cases where the initial decisions are not modified. Finally, columns (6) and (7) present the estimates where slowest 25 percent and fastest 25 percent of evaluators are dropped from the estimation sample. As Table 8 reveals, the estimates are highly consistent across various sub-samples, which indicates that fastness or slowness of decision-making, altering or not altering the initial decisions, or decisions made earlier or later during these six trials have no significant impact on the results. Similarly, there is no appreciable difference between law students and economics students (columns 8 and 9) and between male and female evaluators (columns 10 and 11), with one difference: racial in-group bias is not significantly different from zero in case of female evaluators.[30]

Is it race or something else?

Could these findings be an artifact of some other attribute of the defendants? For example, could it be the case that the body language of the defendants is influenced by their race and that it is the body language, rather than the race of the defendants, that triggers the response of the evaluators? There are a number of answers to this question. First, in our six trials and 2 versions of each trial (generating 12 versions with six white and six minority defendants), body language is very similar between white and minority defendants (see the pictures of the trials with white and minority defendants by scrolling down at http://proficient.ninja/uhasselt/). This is because after shooting a particular trial with the white defendant, that video was shown to the minority defendant (who would replace the white defendant in that particular video) so that he could mimic the body language of the white defendant.

[30] There are 10 female minority evaluators and 9 male minority evaluators.

The body language of the defendants did not project any disrespect towards the judge or the prosecutor either.[31] The only possible difference was the white defendant in trial 1, who occasionally crossed his legs, while other defendants never crossed their legs. If crossing legs while sitting in front of the judge is considered as disrespectful behavior, which should be punished, this would imply that this white defendant received harsher punishment than what was appropriate, and it indicates that our results are potentially an underestimate of the racial bias we detect against minorities.

One can argue that minorities would be more timid and nervous during a trial because of cultural reasons, while white defendants would be more self-confident. Self-confidence could signal innocence, while being timid and nervous could suggest the acceptance of guilt, and this could be the reason why the evaluators were biased against minorities. This argument has no validity because, as discussed above, minority defendants were told to mimic the body language of the white defendants. Second, recall that minority defendants are more likely to get convicted by *both* white and minority evaluators. Given that minority evaluators would not fall into such a trap of cultural misunderstanding of the body language of their own in-group, this result cannot be attributed to possible differences in body language.

VIII. The Analysis of Attorney Decisions

As described in the data section, 36 attorneys also participated in the experiment. Tables 9 and 10 present the descriptive statistics of the attorney sample. It is interesting to note that only 28 percent of lawyers revealed that they had trust in the legal system, when the rate was 43 percent among law and economics students, and 36 percent in the Belgian population (see

[31] The spoken language did not differ between the defendants. They did not speak during the trial with the exception of their statements regarding their understanding of the charges, their statement about a "not guilty" plea and their statement about not adding anything else to their attorneys' defense. These statements were: "Yes," "Not guilty" and "No."

footnote 7). The rate of trust in police is identical between attorneys and students (42 percent), while it is 58 percent in the overall population. The proportion of attorneys who believe that Belgium became a worse place to live because of immigration is 67 percent (58 percent among students, and 75 percent in the population), and 75 percent of the attorneys believe that terrorism is an important problem in Belgium (47 percent among students).

Table 10 displays the summary statistics related to the conviction and sentencing decisions made by attorneys. Lawyers convict at a lower rate in comparison to students (about 10 percentage points lower), but consistent with the behavior of students, lawyers too convict minority defendants at a higher rate (69 percent conviction rate among minority defendants vs 55 percent conviction rate among white defendants). Average effective prison sentence is similar between attorneys and students, and the average fine assigned by attorneys is only slightly lower (420 Euros vs. 468 Euros).[32]

Because there is no racial variation within lawyers (they are all white), in-group bias in lawyer decisions cannot be analyzed. Instead, we focus on estimation of Equations (3A) which investigates the existence of racial bias. Table 11A displays the results obtained from the lawyer sample regarding racial bias in conviction decision, and shows that minority defendants are about 14 percentage points more likely to be found guilty by practicing lawyers, holding constant everything else about the attributes of the case and trial environment. Estimating equation (3B) that includes lawyer fixed effects provided identical coefficient for the *Minority Defendant* dummy. It is interesting to note that the size of the racial bias coefficient estimated in the sample of attorneys (14 percentage points--shown in Table 11A) is almost the same as the one estimated among the sample of law students (14.5 percentage points--shown in Table 8, column 10).

[32] These are weighted averages of fines assigned to minority and white defendants shown in row (7) of tables 3 and 10.

Tables 11B and 11C present the regression results that analyze the impact of defendant race on prison sentence and fine assigned by lawyers.[33] While Table 11B indicates that minority defendants receive sentence lengths that are on average 0.7 month longer, this effect is not statistically different from zero. On the other hand, Table 11C reveals that lawyers assign 131 percent larger fines to convicted minority defendants (implied by the coefficient of 0.927). Thus, the analysis of lawyers' decisions reveal a pattern that is similar to those observed in the behavior of law students and economics students. Minority defendants are more likely to get convicted for the same exact crime, based on the same arguments made by the prosecutors and for the same defense by their attorney, and they are more likely to receive stiffer punishment upon conviction.

Because there are only 36 attorneys in the sample, it is not feasible to divide them into groups (slow vs. fast decision makers, first three decisions vs. last three decisions, and so on) in order to analyze whether the decisions differ between groups. Instead, we added the attorneys to the student sample, and re-ran the various models reported in Table 8. In these models we included an indicator for attorneys and interacted it with the *Minority Defendant* dummy to analyze whether attorney decisions differ from those of the students in various sub-groups. The results are summarized in Table 12. The entries are the coefficients (standard errors) of the *Minority Defendant* dummy in each regression. The coefficient of the interaction term between *Minority Defendant* dummy and the *Attorney* dummy was never different from zero in any of the regressions. The results displayed in Table 12 are very similar to those shown in Table 8,

[33] Each of the 36 lawyers evaluated each of the six cases, yielding to 216 decisions on conviction versus acquittal (108 cases of white defendants and 108 cases of minority defendants). Fifty-nine of the white defendants are convicted (55 percent), while conviction rate among minority defendants was 69 percent (74 minority defendants are convicted). This implies the existence of 15 "excess" convicted minority defendants. Ranking of the 74 convicted minorities by sentence length and fine showed that 12 defendants received zero effective prison term and zero effective fine, despite their conviction. These are the marginal defendants, who are arguably not guilty despite being convicted. Of the reaming convicted defendants with prison terms of zero months, we dropped three people who are sentenced to zero months effective prison term and the lowest initial fines assigned as punishment (two people with 100 Euro fines, and one person with 104 Euros).

indicating that analyzing attorneys and students as one group yields the same results as those obtained from the analysis of students, implying no significant differences in behavior between these two groups of evaluators.

IX. The Impact of Concerns about Immigration, Terrorism, Trust in Police, and Trust in the Legal System

We also analyzed the question of whether the conviction and sentencing decisions of the evaluators get impacted by their trust in the judiciary, or by their trust in police. Similarly, we analyzed whether evaluators' beliefs that terrorism is a major problem in Belgium and whether Belgium is made worse off because of immigration has an impact on the results. In each case we created dichotomous indicators to represent the lack of trust or the concern about terrorism and migration as described in the data section. These variables are added to models (3A) and (3B) and they are interacted with *Minority Defendant, Minority Evaluator* and *Minority Defendant x Minority Evaluator*. These regressions showed that the coefficients of the triple interaction term were almost always statistically insignificant, indicating that evaluators' trust in the judiciary and trust in police, or their concerns about terrorism or migration had no impact on the conviction or sentencing decisions they made with the following exceptions. In the sample of law students and economics students, the concern about immigration increased the racial bias in prison sentencing. Specifically, if the evaluators believe that Belgium is made a worse place to live by people coming to live here from other countries, they assign sentence lengths to minority defendants that are 1.6 months longer in comparison to those who do not believe that immigration has made Belgium a worse place to live. Similarly, those attorneys who do not trust police are about 20 percentage points more likely to convict a minority defendant in comparison to attorneys who have trust in police.

X. Conclusion

In this paper we ask whether the decision about guilt vs. innocence of an individual, who is being accused of a crime, is impacted by his race. We also ask whether the extent of the punishment he receives (prison sentence and fine) depends on his race. Do these judicial decisions depend on whether or not the defendant and the judge are of the same race; i.e. is there in-group bias in judicial decisions? If so, are these influences stronger in case of white judges or minority judges? If there exist race effects on these decisions, are they impacted by whether the judge believes that immigration is a problem, that terrorism is a problem, or whether the judicial system or police can be trusted?

Some of these are old questions, and all of them are important for both scientific inquiry and public policy. These questions, however, have been difficult to answer because of a number of inherent endogeneity issues that emerge in the analysis of data. For instance, even if defendants (some of whom are white and some of whom are minority) are randomly assigned to judges (who also vary in their race), the composition and the attributes of the cases adjudicated by judges are influenced by actors other than the judge and the defendant. As an example, if prosecutors are more lenient and generous towards a particular group of defendants during the plea-bargaining stage, or if prosecutors are less diligent and less aggressive towards a group of individuals during the trial, the conviction and sentencing decisions of judges will resemble favoritism towards that group even if judges themselves are unbiased. Similarly, the quality of the defense attorney may be correlated with the race of the defendants. For example, if minorities are less likely to afford high quality defense attorneys, it will be more likely for minorities to get convicted in a trial. As detailed in Section II, there are other, more subtle, issues that complicate the analyses of judicial decisions and their interplay with race.

At the heart of the issue lies the near impossibility to create a ‘counterfactual scenario’ to a trial, which involves the defense attorney, the prosecutor, and the defendant. In this paper we

create a design which holds constant everything that takes place during a trial, with the exception of the race of the defendant.

We shot Virtual Reality videos of six criminal trials in Belgium, prosecuted by actual prosecutors and defended by actual defense attorneys in an actual courtroom. Only the defendants in the courtroom are actors. The Virtual Reality technology enabled us to replace white defendants in the courtroom with individuals who have Middle Eastern or North African descent. This allowed us to alter only the race of the defendants in these trials, holding all activity in the courtroom constant, including every word spoken by the prosecutor and the defense attorney, and all the body language in the courtroom. A short clip of two videos can be seen here http://proficient.ninja/splitscreen/. This paper is the first one that utilizes 3D Virtual Reality technology that uses actual people, rather than computer-animated scenes.

Scrolling down at the link http://proficient.ninja/uhasselt/ allows one to observe scenes from all six trials. Full versions of one of these trials can be also be watched (in 2D) at this link. Two versions of this particular trial are titled Video No 4 - Full Version 1 (with the minority defendant), and Video No 4, Full Version 2 (with the white defendant).

This design allows us to bypass the identification challenges faced by previous empirical research, and to create arguably perfect counterfactuals. In our case, the race of the defendant in each trial is uncorrelated with the characteristics of the prosecutors, with the characteristics of defense attorneys, and with any activity in the courtroom. Defendant race is also uncorrelated with evaluator attributes. This feature of the design not only permits us to investigate the existence of in-group bias, but it also allows us to decompose the bias to determine its source.

The only variation in a given trial is obtained from the skin color of the defendants. We hold all actions (body language, spoken words, and so on) of the prosecutor and the defense attorney in the courtroom constant. Evaluators who watched the two versions of a particular trial, observed identical courtroom activity with only one difference: half of the evaluators saw a

white defendants, the other half saw a minority defendant. The evaluators don't know the names of the defendants; thus names cannot be used as signals of minority status (Bertrand and Mullainathan 2004). The defendants barely spoke during the trial: they spoke only three times during the trial to answer three questions of the judge about whether they understand the charges against them (they answered "Yes"); whether they plead guilty or not guilty (they answered "Not guilty"), and whether they had anything else to add to what their lawyer said during the trial (they answered "No"). All actor-defendants are born in Belgium and they speak fluent Dutch. Because they had no accent and because they spoke only a few words during the trial their minority status could not be inferred from the way they spoke. Finally, all actor-defendants were instructed to wear similar clothes. Thus, the differentiation between white and minority defendants is obtained from the variation in their skin color. This point is verified by providing pictures of the defendants to another group of 89 freshmen students of Hasselt University. These students identified the race of each defendants correctly with 99 percent accuracy.

A total of 156 Master's degree law students and undergraduate and Master's degree economics students are randomly assigned to watch, from the view point of the judge, the VR videos of these identical criminal trials that differed only in defendants' skin color (See Figures 1 and 2). Background information obtained from the evaluators enabled us to identify their cultural heritage. Evaluators made decisions on guilt/innocence as well as prison sentence and fine in accordance with the guidelines provided by the relevant law.

The results show that minority defendants are more likely to get convicted in comparison to white defendants, even though they are tried for the same exact case, and even though everything that went on during the trial is the same. Further analyses show that this bias is driven equally by white evaluators and minority evaluators. Because minority defendants are more likely to get convicted by *both* white and by minority evaluators, this translates into a racial bias in against minority defendants where they are 12 percentage point more likely to get

convicted in comparison to their white counterparts. This result suggests the existence of statistical discrimination against minorities by minority evaluators, under the assumption that evaluator would not have a distaste towards their own in-group (ruling out taste-based discrimination against one's own in-group). White evaluators' higher propensity to convict minority defendants, on the other hand, may be the result of the combination of both statistical discrimination and taste-based discrimination against minorities

Convicted defendants can be assigned a prison sentence and/or a fine. We find that the defendants receive shorter prison terms if the evaluator is of the same race: convicted defendants receive prison sentences that are about 2.5 months shorter if the evaluator if of the same race. We show that this positive in-group bias in prison sentences is driven by white evaluators. Because most evaluators are white, this behavior translates into racial bias in prison sentencing with minority defendants being sentenced on average 1.3 months longer prison terms, which corresponds to a 32 percent increase in prison time.

The same picture emerges in the assignment of fines: there is in group bias, driven by white evaluators, which translates into racial bias in fines where minority defendants receive fines that are 172 percent higher than their white counterparts. This racial bias in sentencing by white evaluators against minorities is a reflection of taste-based discrimination because statistical discrimination is not relevant at the sentencing stage.

We repeated the same experiment with a group of 36 practicing attorneys, most of whom specialize in criminal law. Because all of these attorney are White, we could not investigate in-group bias in their decisions but analyzed the existence of racial bias. The results were similar to those obtained from law students and economics students, both qualitatively and in magnitude. We find that, in line with the result obtained from students, when attorneys make the adjudication decisions, minority defendants are more likely to get convicted for the same

exact crime, based on the same arguments made by the prosecutors and for the same defense by their attorney, and attorneys assign bigger fines to minority defendants upon conviction.

A large number of additional analyses confirmed the robustness of the results. For example, the speed with which the evaluators made their decisions, or whether the evaluators altered their original conviction or sentencing decisions have no impact on the results. Similarly, the decisions made during the first half of the experiment (the first three trials) are no different from those made in the second half, and the decisions of men and women are similar.

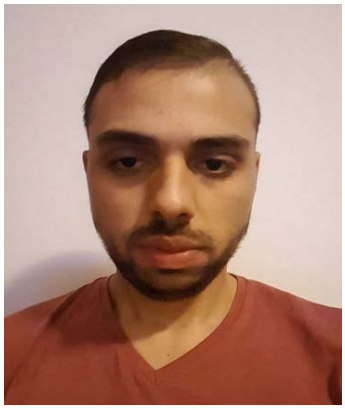
We also analyze whether disparate treatment of defendants is impacted by evaluators' trust in the judiciary, by their trust in police, and by their concerns about terrorism and immigration. We find that with two exceptions[34] evaluators' trust in the judiciary and their trust in police, or their concerns about terrorism or migration had no impact on their conviction or sentencing decisions, suggesting that the source of the racial bias may be deep-rooted.

[34] Law and economics students' concern about immigration lead to longer prison terms for minorities, and lawyers' lack of trust in police lead to higher probability of conviction of minorities.

Figure 1
A Snapshot of the Virtual Reality Videos of the same Trial

This scene can be watched at http://proficient.ninja/splitscreen/

Figure 2
Six Defendants in Six Trials



Minority Defendant-1 (M1)

White Defendant-1 (WH1)

Minority Defendant-2 (M2)

White Defendant-2 (WH2)

Minority Defendant-3 (M3)

White Defendant-3 (WH3)

Figure 3
Evaluators Watching the VR Videos of the Trials

Figure 4
Distribution of Time Taken to Make Decisions (in Seconds)

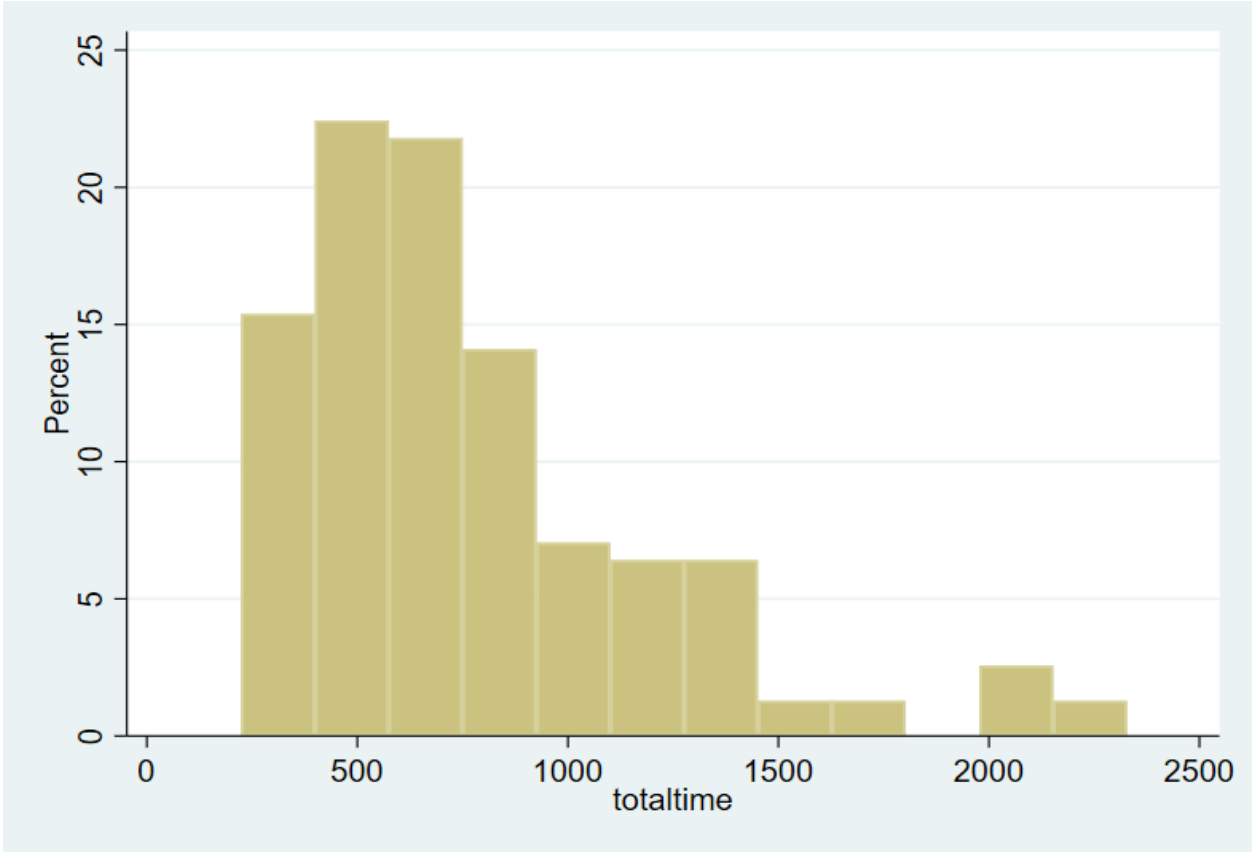


Table 1

The Sequence of Trials Watched by Evaluators, and the Identity of Defendants, Prosecutors an Defense Attorneys in Each Trial

Set 1 (Sequence1, Version 1)	Set 2 (Sequence1, Version 2)	Set 3 (Sequence2, Version 1)	Set 4 (Sequence2, Version 2)	Set 5 (Sequence3, Version 1)	Set 6 (Sequence3, Version 2)
$^{WH1}_{SO}BUR^{1}_{B}$	$^{M1}_{SO}BUR^{1}_{B}$	$^{M2}_{TO}AS^{3}_{P}$	$^{WH2}_{TO}AS^{3}_{P}$	$^{WH2}_{TO}BUR^{3}_{B}$	$^{M2}_{TO}BUR^{3}_{B}$
$^{M1}_{SO}BUR^{2}_{B}$	$^{WH1}_{SO}BUR^{2}_{B}$	$^{WH3}_{M}AS^{2}_{P}$	$^{M3}_{M}AS^{2}_{P}$	$^{M3}_{M}AS^{1}_{P}$	$^{WH3}_{M}AS^{1}_{P}$
$^{WH2}_{TO}BUR^{3}_{B}$	$^{M2}_{TO}BUR^{3}_{B}$	$^{M3}_{M}AS^{1}_{P}$	$^{WH3}_{M}AS^{1}_{P}$	$^{M1}_{SO}BUR^{2}_{B}$	BUR^{2}_{B}
$^{M3}_{M}AS^{1}_{P}$	$^{WH3}_{M}AS^{1}_{P}$	$^{WH2}_{TO}BUR^{3}_{B}$	$^{M2}_{TO}BUR^{3}_{B}$	$^{WH3}_{M}AS^{2}_{P}$	$^{M3}_{M}AS^{2}_{P}$
$^{WH3}_{M}AS^{2}_{P}$	$^{M3}_{M}AS^{2}_{P}$	$^{M1}_{SO}BUR^{2}_{B}$	$^{WH1}_{SO}BUR^{2}_{B}$	$^{WH1}_{SO}BUR^{1}_{B}$	$^{M1}_{SO}BUR^{1}_{B}$
$^{M2}_{TO}AS^{3}_{P}$	$^{WH2}_{TO}AS^{3}_{P}$	$^{WH1}_{SO}BUR^{1}_{B}$	$^{M1}_{SO}BUR^{1}_{B}$	$^{M2}_{TO}AS^{3}_{P}$	$^{WH2}_{TO}AS^{3}_{P}$

Each evaluator was randomly assigned to one of these six sets. Each set includes six cases (trials), that are watched in the order listed in each column. Each cell, such as $^{WH1}_{SO}BUR^{1}_{B}$, summarizes the characteristics of that particular trial. BUR and AS stand for a burglary case, and an assault case, respectively. BUR^{1} means the first burglary case, BUR^{2} means the second burglary case, AS^{3} stands for the third assault case, and so on.

A subscript to the right identifies the prosecutor: TH^{1}_{B} means that the first burglary case is prosecuted by Bruno (B). There are two prosecutors in the experiment: Bruno (B) and Pieter (P). The subscript to the left identifies the defense attorney. There are three defense attorneys: TO, SO and M.

The superscript to the left identifies the defendant. There are three minority defendants: M1, M2, and M3; and there are three white defendants: WH1, WH2 and WH3. See Figure 2 for the pictures of all six defendants.

The entry $^{WH1}_{SO}BUR^{1}_{B}$ represents the first burglary case (BUR^{1}), where the defendant was WH1. The case was prosecuted by B, and the defense attorney was SO.

Note that the first VR video of Set 1 is $^{WH1}_{SO}BUR^{1}_{B}$ and the first video of Set 2 is $^{M1}_{SO}BUR^{1}_{B}$. These two videos are identical in all respects but one: the race of the defendant (WH1 vs. M1); See Figure 1.

Table 2
Descriptive Statistics of Evaluator Characteristics

	Evaluator Attributes
Minority Evaluator	0.12 (0.33)
Male Evaluator	0.47 (0.50)
Law Student	0.43 (0.50)
N	156
	Evaluator Beliefs
Trust in the Legal System	0.43 (0.50)
N	154
Trust in Police	0.42 (0.49)
Terrorism is an Important Problem in Belgium	0.47 (0.50)
Belgium is made a worse place to live by people coming to live here from other countries	0.58 (0.49)
N	156

Table 3
Descriptive Statistics of Conviction Decision, Sentence Length and Fine Imposed by Evaluators

			White Evaluator		Minority Evaluator	
	Minority defendant (1)	White defendant (2)	Minority Defendant (3)	White Defendant (4)	Minority Defendant (5)	White Defendant (6)
(1) Conviction Rate	**0.79*** (0.41)**	**0.68*** (0.47)**	**0.78*** (0.41)**	**0.68*** (0.47)**	**0.84** (0.37)**	**0.67** (0.48)**
N	468	468	411	411	57	57
(2) Initial Prison Sentence	10.2 (8.6)	9.7 (7.9)	10.5 (8.9)	9.5 (7.4)	8.2* (6.0)	11.3* (10.6)
N	370	317	322	279	48	38
(3) Suspended Prison Sentence	5.8 (5.2)	5.8 (5.6)	5.9 (5.2)	5.9 (5.6)	4.8 (4.5)	5.0 (5.4)
N	370	317	322	279	48	38
(4) Effective Prison Sentence	**4.4 (6.7)**	**3.9 (5.3)**	**4.6** (7.1)**	**3.6** (4.9)**	**3.3** (3.8)**	**6.3** (7.3)**
N	370	317	322	279	48	38
(5) Initial Fine	792 (1,213)	703 (1,092)	789 (1,248)	681 (1,097)	808 (957)	868 (1,057)
N	370	317	322	279	48	38
(6) Suspended Fine	290 (555)	275 (516)	289 (556)	271 (526)	290 (553)	305 (442)
N	370	317	322	279	48	38
(7) Effective Fine	**502 (1,019)**	**428 (895)**	**500 (1,054)**	**409 (903)**	**517 (755)**	**562 (837)**
N	370	317	322	279	48	38

* signifies difference in the means between groups at the 10% level. ** indicates significant difference at the 5% level, and *** represents difference at the 1* level or better.

Table 4A
In-group Bias in Conviction Decisions

	(1)	(2)	(3)	(4)
Minority Defendant	0.102***	0.110***	0.110***	0.110***
	(0.029)	(0.025)	(0.026)	(0.025)
Minority Evaluator	-0.011	-0.009	-0.017	-0.016
	(0.078)	(0.073)	(0.072)	(0.072)
Minority Defendant x Minority Evaluator	0.074	0.071	0.073	0.071
	(0.094)	(0.078)	(0.081)	(0.078)
Law Student	0.011	0.011	-0.017	-0.018
	(0.029)	(0.029)	(0.029)	(0.029)
Male Evaluator			-0.093***	-0.100***
			(0.029)	(0.029)
Early Trial	-0.123***		-0.098***	
	(0.028)		(0.030)	
Sequence 1				-0.074**
				(0.034)
Sequence 2				-0.042
				(0.031)
Assault Case	-0.106***			
	(0.029)			
Observations	936	936	936	936
Trial FE	No	Yes	Yes	Yes

Early Trial is a dummy that takes the value of 1 if the trial is one of the first three trials watched by that evaluator (as opposed to the last three). Sequence 1 and Sequence 2 are dichotomous indicators of the order in which the videos are watched (see table 1). Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 4B
In-group Bias in Conviction Decisions -
Models with Evaluator Fixed Effects

	(1)	(2)	(3)
Minority Defendant	0.104***	0.102***	0.110***
	(0.027)	(0.029)	(0.026)
Minority Defendant X Minority Evaluator	0.074	0.074	0.073
	(0.089)	(0.094)	(0.081)
Early Trial	-0.113***	-0.123***	-0.098***
	(0.029)	(0.028)	(0.030)
Assault Case		-0.106***	
		(0.029)	
Observations	936	936	936
Number of Evaluators	156	156	156
Evaluator FE	Yes	Yes	Yes
Trial FE	No	No	Yes

Early Trial is a dummy that takes the value of 1 if the trial is one of the first three trials watched by that evaluator (as opposed to the last three). Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 4C
Decomposing In-group Bias in Conviction Decisions:
Conviction Regressions Conditional on Defendant Race

	(1)	(2)
White Evaluator	0.026	
	(0.072)	
Minority Evaluator		0.065
		(0.054)
Law Student	-0.052	0.018
	(0.044)	(0.036)
Male Evaluator	-0.111**	-0.076**
	(0.043)	(0.037)
Early Trial	-0.089*	-0.109***
	(0.046)	(0.040)
Observations	468	468
Sample	White Defendants	Minority Defendants
Trial FE	Yes	Yes

Early Trial is a dummy that takes the value of 1 if the trial is one of the first three trials watched by that evaluator (as opposed to the last three). Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 4D
Racial Bias in Conviction Decisions

	(1)	(2)	(3)	(4)
Minority Defendant	0.111***	0.119***	0.119***	0.119***
	(0.028)	(0.024)	(0.024)	(0.024)
Law Student	0.008	0.008	-0.019	-0.020
	(0.028)	(0.028)	(0.028)	(0.028)
Male Evaluator			-0.094***	-0.101***
			(0.028)	(0.029)
Early Trial	-0.123***		-0.098***	
	(0.028)		(0.030)	
Sequence 1				-0.074**
				(0.034)
Sequence 2				-0.042
				(0.031)
Assault Case	-0.106***			
	(0.029)			
Observations	936	936	936	936
Trial FE	No	Yes	Yes	Yes

Early Trial is a dummy that takes the value of 1 if the trial is one of the first three trials watched by that evaluator (as opposed to the last three). Sequence 1 and Sequence 2 are dichotomous indicators of the order in which the videos are watched (see Table 1).Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 4E
Racial Bias in Conviction Decisions
Models with Evaluator Fixed Effects

	(1)	(2)	(3)
Minority Defendant	0.113***	0.111***	0.119***
	(0.026)	(0.028)	(0.024)
Early Trial	-0.113***	-0.123***	-0.098***
	(0.029)	(0.028)	(0.030)
Assault Case		-0.106***	
		(0.029)	
Observations	936	936	936
Number of Evaluators	156	156	156
Evaluator FE	Yes	Yes	Yes
Trial FE	No	No	Yes

Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 5
The Conviction Rate and the Length of Effective Prison Sentence
by Defendant and Evaluator Race

	White Evaluator		Minority Evaluator	
	Minority Defendant (1)	White Defendant (2)	Minority Defendant (3)	White Defendant (4)
Conviction Rate	0.78***	0.68***	0.84**	0.67**
	(0.41)	(0.47)	(0.37)	(0.48)
N	411	411	57	57
Effective Prison Sentence (in months)	4.61**	3.60**	3.33**	6.26**
	(7.06)	(4.91)	(3.81)	(7.31)
N	322	279	48	38

White Evaluators made 411 conviction decisions on minority defendants and 411 conviction decisions on white defendants (columns 1 and 2). Minority Evaluators made 57 conviction decisions on each group (columns 3 and 4). Effective prison sentence is assigned by the evaluators on those who are convicted (found guilty). Mean-comparison test: *** p<0.01, ** p<0.05, * p<0.1

Table 6A
In-group Bias in Prison Sentencing
Dependent Variable: Effective Prison Sentence (months)

A: Regressions using trimmed sample				
	(1)	(2)	(3)	(4)
Minority Defendant	1.703***	1.768***	1.765***	1.750***
	(0.554)	(0.545)	(0.541)	(0.552)
Minority Evaluator	2.542*	2.527*	2.746**	2.684**
	(1.368)	(1.354)	(1.292)	(1.310)
Minority Defendant	-3.579***	-3.647***	-3.728***	-3.783***
x Minority Evaluator	(1.103)	(1.138)	(1.150)	(1.197)
Law Student	-0.956	-1.014	-0.462	-0.444
	(0.664)	(0.659)	(0.668)	(0.674)
Observations	634	634	634	634
Trial FE	No	Yes	Yes	Yes
B: Regressions using untrimmed sample				
	(5)	(6)	(7)	(8)
Minority Defendant	1.083**	1.198**	1.193**	1.185**
	(0.508)	(0.492)	(0.491)	(0.492)
Minority Evaluator	2.538*	2.537*	2.747**	2.710**
	(1.367)	(1.356)	(1.295)	(1.308)
Minority Defendant	-3.784***	-3.886***	-3.924***	-3.989***
x Minority Evaluator	(1.129)	(1.153)	(1.150)	(1.194)
Law Student	-0.956	-0.963	-0.440	-0.406
	(0.628)	(0.626)	(0.633)	(0.639)
Observations	687	687	687	687
Trial FE	No	Yes	Yes	Yes

Effective Prison Sentence is equal to initial prison sentence minus suspended sentence. Models in column (1) include Early Trial and Assault Case dummies. Models in column (2) include Trial Fixed Effects. Models in column (3) include Trial Fixed Effects, Male Evaluator and Early Trial. Models in column (4) include Trial Fixed Effects, Male Evaluator and Sequence dummies. Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 6B
Decomposing In-group Bias in Prison Sentencing
Effective Prison Sentence Regressions Conditional on Defendant Race

	(1)	(2)	(3)	(4)
White Evaluator	-2.778**		-2.778**	
	(1.329)		(1.329)	
Minority Evaluator		-1.047		-1.247
		(0.975)		(0.861)
Law Student	-0.212	-0.812	-0.212	-0.685
	(0.696)	(0.934)	(0.696)	(0.818)
Male Evaluator	1.318*	1.940*	1.318*	1.842**
	(0.707)	(0.984)	(0.707)	(0.882)
Early Trial	-1.251**	-0.773	-1.251**	-0.811
	(0.563)	(1.021)	(0.563)	(0.845)
Observations	317	317	317	370
Sample	White Defendants	Minority Defendants	White Defendants	Minority Defendants
Trial FE	Yes	Yes	Yes	Yes
Trimmed Sample	Yes	Yes	No	No

Early Trial is a dummy that takes the value of 1 if the trial is one of the first three trials watched by that evaluator (as opposed to the last three). Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 6C
Racial Bias in Prison Sentencing:
Dependent Variable: Effective Prison Sentence (months)

A: Regressions using trimmed sample				
	(1)	(2)	(3)	(4)
Minority Defendant	1.276**	1.331***	1.320***	1.298**
	(0.509)	(0.504)	(0.497)	(0.509)
Law Student	-1.054	-1.107*	-0.593	-0.563
	(0.652)	(0.650)	(0.655)	(0.660)
Male Evaluator			1.670**	1.728**
			(0.713)	(0.752)
Observations	634	634	634	634
B: Regressions using untrimmed sample				
	(5)	(6)	(7)	(8)
Minority Defendant	0.619	0.720	0.712	0.695
	(0.468)	(0.456)	(0.452)	(0.455)
Law Student	-1.017	-1.019	-0.532	-0.488
	(0.617)	(0.617)	(0.621)	(0.626)
Male Evaluator			1.621**	1.642**
			(0.678)	(0.709)
Observations	687	687	687	687

Models in column (1) include Early Trial and Assault Case dummies. Models in column (2) include Trial Fixed Effects. Models in column (3) include Trial Fixed Effects and Early Trial. Models in column (4) include Trial Fixed Effects and Sequence dummies. Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 7A
In-group Bias in Fines
Dependent Variable: Effective Fine (log Euros)

A: Regressions using trimmed sample				
	(1)	(2)	(3)	(4)
Minority Defendant	1.169***	1.143***	1.140***	1.143***
	(0.209)	(0.206)	(0.205)	(0.207)
Minority Evaluator	1.336**	1.375***	1.367***	1.390***
	(0.545)	(0.518)	(0.523)	(0.518)
Minority Defendant	-1.078*	-1.014*	-1.042*	-1.017*
x Minority Evaluator	(0.550)	(0.549)	(0.543)	(0.546)
Law Student	-0.673**	-0.626**	-0.592**	-0.599**
	(0.273)	(0.264)	(0.283)	(0.281)
Observations	634	634	634	634
Trial FE	No	Yes	Yes	Yes
B: Regressions using untrimmed sample				
	(5)	(6)	(7)	(8)
Minority Defendant	0.573***	0.535**	0.538**	0.535**
	(0.210)	(0.209)	(0.208)	(0.209)
Minority Evaluator	1.341**	1.381***	1.376***	1.419***
	(0.545)	(0.519)	(0.523)	(0.526)
Minority Defendant	-1.126**	-1.034*	-1.073*	-1.032*
xMinority Evaluator	(0.559)	(0.563)	(0.568)	(0.564)
Law Student	-0.673**	-0.662**	-0.614**	-0.626**
	(0.296)	(0.290)	(0.307)	(0.305)
Observations	687	687	687	687
Trial FE	No	Yes	Yes	Yes

Models in column (1) include Early Trial and Assault Case dummies. Models in column (2) include Trial Fixed Effects. Models in column (3) include Trial Fixed Effects, Male Evaluator and Early Trial. Models in column (4) include Trial Fixed Effects, Male Evaluator and Sequence dummies. Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 7B
Decomposing In-group Bias in Fines
Effective Fine Regressions Conditional on Defendant Race
Dependent variable: Log (Euros)

	(1)	(2)	(3)	(4)
White Evaluator	-1.387***		-1.387***	
	(0.520)		(0.520)	
Minority Evaluator		0.295		0.278
		(0.421)		(0.508)
Law Student	-0.550	-0.673**	-0.550	-0.684**
	(0.386)	(0.300)	(0.386)	(0.346)
Male Evaluator	0.216	-0.092	0.216	0.008
	(0.385)	(0.287)	(0.385)	(0.342)
Early Trial	0.522	0.822***	0.522	0.785***
	(0.346)	(0.294)	(0.346)	(0.284)
Observations	317	317	317	370
Sample	White Defendants	Minority Defendants	White Defendants	Minority Defendants
Trial FE	Yes	Yes	Yes	Yes
Trimmed Sample	Yes	Yes	No	No

Early Trial is a dummy that takes the value of 1 if the trial is one of the first three trials watched by that evaluator (as opposed to the last three). Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 7C
Racial Bias in Fines
Dependent Variable: Effective Fine (log Euros)

A: Regressions using trimmed sample				
	(1)	(2)	(3)	(4)
Minority Defendant	1.041***	1.022***	1.016***	1.023***
	(0.197)	(0.193)	(0.192)	(0.194)
Law Student	-0.763***	-0.723***	-0.699**	-0.709**
	(0.265)	(0.258)	(0.273)	(0.273)
Male Evaluator			0.087	0.056
			(0.275)	(0.271)
Observations	634	634	634	634
B: Regressions using untrimmed sample				
	(5)	(6)	(7)	(8)
Minority Defendant	0.441**	0.415**	0.413**	0.416**
	(0.198)	(0.196)	(0.196)	(0.197)
Law Student	-0.758***	-0.757***	-0.717**	-0.738**
	(0.286)	(0.282)	(0.295)	(0.295)
Male Evaluator			0.096	0.038
			(0.296)	(0.292)
Observations	687	687	687	687

Models in column (1) include Early Trial and Assault Case dummies. Models in column (2) include Trial Fixed Effects. Models in column (3) include Trial Fixed Effects and Early Trial. Models in column (4) include Trial Fixed Effects and Sequence dummies. Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 8
Results of Sensitivity Analyses

	Whole Sample (1)	First 3 decisions (2)	Last 3 decisions (3)	Change (4)	No change (5)	Drop Slow 25% (6)	Drop Fast 25% (7)	Males (8)	Females (9)	Law Students (10)	Econ Students (11)
Racial bias in conviction	0.119***	0.111***	0.126***	0.058*	0.164***	0.129***	0.109***	0.130***	0.112***	0.146***	0.097***
	(0.024)	(0.040)	(0.033)	(0.031)	(0.034)	(0.026)	(0.026)	(0.039)	(0.029)	(0.037)	(0.032)
In-group bias in conviction	0.071	0.108	0.038	-0.029	0.204	0.132***	-0.002	0.152	0.002	0.199	0.051
	(0.078)	(0.107)	(0.105)	(0.068)	(0.126)	(0.028)	(0.086)	(0.121)	(0.100)	(0.142)	(0.092)
Racial bias in prison sentence	1.298**	1.209	1.415**	0.883*	1.676**	1.336**	1.266**	1.616*	0.926**	1.013**	1.506*
	(0.509)	(0.891)	(0.567)	(0.519)	(0.827)	(0.608)	(0.551)	(0.929)	(0.459)	(0.480)	(0.820)
In-group bias in prison sentence	-3.783***	-2.727**	-4.745**	-3.356**	-4.316**	-3.711***	-3.513***	-6.566***	-1.700	-2.529**	-4.209***
	(1.197)	(1.375)	(2.045)	(1.472)	(1.897)	(1.354)	(1.307)	(1.906)	(1.388)	(1.066)	(1.556)
Racial bias in fine	1.023**	1.020***	1.017***	1.038***	1.008***	1.067***	1.011***	1.005***	1.172***	1.105***	0.965***
	(0.194)	(0.298)	(0.267)	(0.271)	(0.278)	(0.214)	(0.212)	(0.291)	(0.245)	(0.294)	(0.260)
In-group bias in fine	-1.017*	-0.432	-1.841**	-1.402*	-0.549	-0.818	-1.075*	-2.016***	-0.102	-0.942	-1.004
	(0.546)	(0.519)	(0.850)	(0.763)	(0.759)	(0.624)	(0.585)	(0.598)	(0.766)	(0.826)	(0.689)

The column Whole Sample re-produces the results reported in Tables 4C, 4A, 6C, 6A, 7C, and 7A. Change stands for the sample of observations related to the cases where the evaluator has made at least one change after making his/her initial decisions on conviction, prison term and fine. No Change represents the cases for which the evaluator did not alter his/her initial decisions. Drop Slow includes the sample that is created after dropping the slowest 25 % of the evaluators based on total time spent in decision making. Drop Fast excludes the 25% fastest evaluators (see Figure 4).

Table 9
Descriptive Statistics of Evaluator Characteristics (Lawyers)

	Evaluator Attributes
Male Evaluator	0.58 (0.50)
N	36
	Evaluator Beliefs
Trust in the Legal System	0.28 (0.45)
Trust in Police	0.42 (0.50)
Terrorism is an Important Problem in Belgium	0.75 (0.44)
N	36
Belgium is made a worse place to live by people coming to live here from other countries	0.67 (0.48)
N	27

Table 10
Descriptive Statistics of Conviction Decision,
Sentence Length and Fine Imposed by Evaluators (Lawyers)

	Minority Defendant (1)	White Defendant (2)
(1) Conviction Rate	**0.69** **(0.47)**	**0.55**** **(0.50)**
N	108	108
(2) Initial Prison Sentence	11.2 (6.5)	11.1 (7.5)
N	74	59
(3) Suspended Prison Sentence	6.2 (4.2)	6.5 (4.1)
N	74	59
(4) Effective Prison Sentence	**5.0** **(6.5)**	**4.5** **(7.5)**
N	74	59
(5) Initial Fine	711 (840)	701 (779)
N	74	59
(6) Suspended Fine	286 (460)	284 (364)
N	74	59
(7) Effective Fine	**425** **(554)**	**417** **(577)**
N	74	59

* signifies difference in the means between groups at the 10% level. ** indicates significant difference at the 5% level, and *** represents difference at the 1* level or better.

Table 11A
Lawyers
Conviction Decisions

	(1)	(2)	(3)	(4)
Minority Defendant	0.139*	0.139**	0.139**	0.139**
	(0.075)	(0.061)	(0.060)	(0.061)
Male Evaluator			-0.106*	-0.108*
			(0.061)	(0.055)
Early Trial	-0.123**		-0.104	
	(0.056)		(0.063)	
Sequence 1				0.008
				(0.068)
Sequence 2				-0.149**
				(0.061)
Assault Case	-0.190***			
	(0.061)			
Observations	216	216	216	216
Trial FE	No	Yes	Yes	Yes

Early Trial is a dummy that takes the value of 1 if the trial is one of the first three trials watched by that evaluator (as opposed to the last three). Sequence 1 and Sequence 2 are dichotomous indicators of the order in which the videos are watched (see table 1). Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 11B
Lawyers
Bias in Prison Sentencing
Dependent Variable: Effective Prison Sentence (months)

	(1)	(2)	(3)	(4)
Minority Defendant	0.896	0.683	0.707	0.716
	(0.835)	(0.882)	(0.890)	(0.862)
Male Evaluator			2.138	1.840
			(1.717)	(1.425)
Early Trial	-1.171		-0.298	
	(1.311)		(1.379)	
Sequence 1				1.241
				(2.447)
Sequence 2				0.202
				(1.022)
Assault Case	-5.495***			
	(0.766)			
Observations	118	118	118	118
Trial FE	No	Yes	Yes	Yes

Early Trial is a dummy that takes the value of 1 if the trial is one of the first three trials watched by that evaluator (as opposed to the last three). Sequence 1 and Sequence 2 are dichotomous indicators of the order in which the videos are watched (see table 1). Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 11C
Lawyers
Bias in Fines
Dependent Variable: Effective Fine (log Euros)

	(1)	(2)	(3)	(4)
Minority Defendant	0.941**	0.923**	0.925**	0.927**
	(0.436)	(0.443)	(0.430)	(0.421)
Male Evaluator			0.569	0.431
			(0.656)	(0.644)
Early Trial	0.245		0.167	
	(0.413)		(0.457)	
Sequence 1				0.648
				(0.758)
Sequence 2				0.246
				(0.897)
Assault Case	-0.971**			
	(0.409)			
Observations	118	118	118	118
Trial FE	No	Yes	Yes	Yes

Early Trial is a dummy that takes the value of 1 if the trial is one of the first three trials watched by that evaluator (as opposed to the last three). Sequence 1 and Sequence 2 are dichotomous indicators of the order in which the videos are watched (see table 1). Robust standard errors are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Table 12
Sensitivity Analysis for the Entire Sample of Law Students, Economics Students and Practicing Attorneys

	Whole sample (1)	First 3 decisions (2)	Last 3 decisions (3)	Change (4)	No change (5)	Drop Slow 25% (6)	Drop Fast 25% (7)
Overall bias in conviction	0.123***	0.121***	0.125***	0.059*	0.166***	0.131***	0.094***
	(0.023)	(0.036)	(0.031)	(0.030)	(0.031)	(0.027)	(0.025)
	N=1,152	N=576	N=576	N=456	N=696	N=864	N=864
Overall bias in prison sentence	1.266**	1.051	1.484***	1.174**	1.389**	1.182**	0.784**
	(0.442)	(0.718)	(0.511)	(0.471)	(0.701)	(0.595)	(0.315)
	N=752	N=349	N=403	N=325	N=427	N=562	N=589
Overall bias in fine	1.015**	0.903***	1.088***	1.159***	0.898***	0.950***	1.029***
	(0.178)	(0.272)	(0.235)	(0.265)	(0.239)	(0.199)	(0.207)
	N=752	N=349	N=403	N=325	N=427	N=562	N=589

Change stands for the sample of observations related to the cases where the evaluator has made at least one change after making his/her initial decisions on conviction, prison term and fine. No Change represents the cases for which the evaluator did not alter his/her initial decisions. Drop Slow includes the sample that is created after dropping the slowest 25 % of the evaluators based on total time spent in decision making. Drop Fast excludes the 25% fastest evaluators (see Figure 4).

In prison time and fine regressions, the difference in the sample sizes pertaining to the first three and last three decisions is because the conviction rates are lower during the first three decisions. This is also depicted by the negative and significant coefficient of the "Early Trial" dummy in Tables 4A-4D. The difference in the sample sizes between the samples when we drop the slowest and fastest 25 percent of evaluators in prison sentence and fine regressions is because those evaluators who are not dropped contribute different number of observations to the estimation sample.

References

Abrams, David S, Marianne Bertrand and Sendhil Mullainathan. 2012. "Do Judges Vary in Their Treatment of Race?" *The Journal of Legal Studies*, 41:2, 347-383

Alesina, Alberto and Eliana La Ferrara. 2014. "A Test of Racial Bias in Capital Sentencing." *The American Economic Review*, 104:11, 3397-3433

Agan, Amanda, Matthew Freedman and Emily Owens.2018. "Is Your Lawyer a Lemon?Incentives and Selection in the Public Provision of Criminal Defense" NBER Working Paper No. 24579.

Anderson, James M and Paul Heaton. 2012. "How Much Difference Does the Lawyer Make? The Effect of Defense Counsel on Murder Case Outcomes." *Yale Law Journal*, 122:154, 154-217

Angrist, Joshua, Eric Bettinger and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *The American Economic Review*, 96:3, 847-862

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

Argys, Laura M and H Naci Mocan. 2004. "Who Shall Live and Who Shall Die? An Analysis of Prisoners on Death Row in the United States." *The Journal of Legal Studies*, 33:2, 255-282

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94 (4): 991-1013.

Boylan, Richard T. 2012. "The Effect of Punishment Severity on Plea Bargaining." *The Journal of Law and Economics*, 55:3, 565-591

Chen, Daniel, Yosh Halberstam and Alan Yu. 2018 "Covering: Mutable Characteristics and Perceptions of Voice in the U.S. Supreme Court," Toulouse Schhol of Economics Working Paper.

Correll, Joshua, Bernadette Park, Charles Judd and Bernd Wittenbrink. 2002. "The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals." *Journal of Personality and Social Psychology, 83*:6, 1314-1329

Correll, Joshua, Bernadette Park, Charles Judd, Bernd Wittenbrink, Melody Sadler and Tracie Keesee. 2007. "Across the thin blue line: Police officers and racial bias in the decision to shoot." *Journal of Personality and Social Psychology, 92*:6, 1006-1023.

Depew, Briggs, Ozkan Eren and Naci Mocan. 2017. "Judges, Juveniles and In-group Bias." *The Journal of Law and Economics*, 60:2, 209-239

Eren, Ozkan and Naci Mocan. 2018. "Emotional Judges and Unlucky Juveniles." *American Economic Journal: Applied Economics*, 10:3, 171-205

Gazal-Ayal, Oren and Raanan Sulitzeanu-Kenan. 2010. "Let My People Go: Ethnic In-Group Bias in Judicial Decisions—Evidence from a Randomized Natural Experiment." *Journal of Empirical Legal Studies*, 7:3, 403-428

Glaeser, Edward L and Bruce Sacerdote. 2003. "Sentencing in Homicide Cases and the Role of Vengeance." *The Journal of Legal Studies*, 32:2, 363-382

Goette, Lorenz, David Huffman and Stephan Meier. 2006. "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence using Random Assignment to Real Social Groups." *American Economic Review*, 96:2, 212-216

Grossman, Guy, Oren Gazal-Ayal, Samuel Pimentel and jeremy Weinstein, 2016. "Descriptive Representation and Judicial Outcomes in Multiethnic Societies." *American Journal of Political Science*. 60: 1, pp. 44-69.

Hoffman, Florian and Philip Oreopoulos. 2009. "A Professor Like Me: Influence of Professor Gender on University Achievement." *Journal of Human Resources*. 44: 479-494.

Kennedy, Peter E. 1981. "Estimation with Correctly Interpreted Dummy varaibles in Semilogarithmic Equations." *American Economic Review*. p. 801

List, John and Michael Price.2009. " The role of social connections in charitable fundraising: Evidence from a natural field experiment." J*ournal of Economic Behavior and Organization*, 69:2, 160-169

Mendoza, Saaid A, Sean P Lane and David M Amodio. 2014. "For Members Only: Ingroup Punishment of Fairness Norm Violations in the Ultimatum Game." *Social Psychological and Personality Science*, 5:6, 662-670

Mustard, David B. 2001. "Racial, Ethnic, and Gender Disparities in Sentencing: Evidence From the US Federal Courts." *Journal of Law and Economics*, 44:1, 285-314.

Omi, Michael and Howard Winant, 1994. *Racial Formation in the United States*. New York: Routledge.

Parsons, C. A., J. Sulaeman, M. C. Yates, and D. S. Hamermesh (2011). Strike Three: Discrimination, Incentives, and Evaluation. *American Economic Review* 101(4), 1410-35.

Price, Joe and Justin Wolfers (2010). Racial Discrimination among NBA Referees. *Quarterly Journal of Economics* 125(4), 1859-1887.

Rehavi, M Marit and Sonja B Starr. 2014. "Racial Disparity in Federal Criminal Sentences." *Journal of Political Economy*, 122:6, 1320-1354

Sadowski, W., & Stanney, K.M. (2002). Presence in virtual environments. In K. M. Stanney (Eds.). *Handbook of virtual environments: Design, implementation and applications* (pp. 791-806). Mahwah : IEA

Schulman, Kevin A., Jesse A. Berlin, William Harless, Jon F. Kerner, Shyrl Sistrunk, Bernard J. Gersh, D. Phil, Ross Dubé, Christopher K. Taleghani, Jennifer E. Burke, Sankey Williams, John M. Eisenberg, William Ayeand José J. Escarce. 1999, "The Effect of Race and Sex on Physicians' Recommendations for Cardiac Catheterization." *New England Journal of Medicine*, 340, 618-626.

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

Shayo, Moses and Asaf Zussman. 2017. "Conflict and the Persistence of Ethnic Bias." *American Economic Journal: Applied Economics*, 9:4, 137-65.

Slater, M., & Wilbur, S. 1997. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence : Teleoperators and Virtual Environments*, *6*(6), 603-616.

Appendix 1

The Design and Organization of the VR experiment

1. Design of the experiment

We were granted permission by the district attorney of Limburg (Belgium) to use actual criminal cases as well as actual prosecutors and an actual court room.

1.1.Selection of cases

We had an initial discussion with two Belgian prosecutors about the case types which could be suitable for this research purpose. The criteria that were discussed with the prosecutors during the selection process of the case types were:

- the range of the potential sentences allowed by the penal code for a given case type (e.g. we are less interested in cases where the discretionary powers of the judge are limited);
- the complexity of the general case type (e.g. no medical liability cases in which one has to study numerous medical documents to understand the nature of the case)

Burglaries and assaults fit these criteria. Thus, we asked the prosecutors who specialized in either burglaries or assault to each select seven actual cases from the archive (or the docket) so that we could review them. We asked them to consider cases in which the defendant pleaded not guilty in order to avoid clear cut outcomes. After elaborate discussions with the prosecutors on the content of each of these cases, we kept three burglary and three assault case files that were suited for the purpose of this research.

We did not change any actual facts of the cases. The names of the actual victims, however, have been changed to guarantee anonymity of the real cases. We do not use names for the defendants because the names need to fit for both the white and non-white defendant. For the burglaries we changed the location of the facts as well. Given that the actual defendants were in some cases older than our actor-defendants, we changed the age to make it more realistic. Otherwise, no substantive changes were made to the content of the cases.

1.2.Selection of the court room

We decided to organize the shooting of all the videos in one court room (keeping it constant over all videos). We reviewed three court rooms in the main court building in Hasselt (Belgium) and decided to use a small court room (which is better for purposes of the 360 degree camera). This court room is in fact also being used for burglaries and assault cases.

1.3.Selection of prosecutors, lawyers and the judge

The district attorney informed all prosecutors of the region of Limburg that we were looking for two prosecutors to participate in the experiment. Two prosecutors experienced in (among others) burglary and assault cases and with camera experience volunteered to participate in the experiment.

We also contacted a number of law firms with experience in criminal cases to participate in the experiment. We had five actual criminal lawyers that were eager to participate and after an interview we selected three lawyers (two female and one male) with actual experience in both burglary and assault cases. The lawyers were given the case files one week before the shooting of the videos and we asked them to prepare it similar to an actual court hearing. The lawyers did not talk to the prosecutors before the shooting, in order to keep the courtroom interactions as realistic as possible.

Given the nature of proceedings in criminal cases in Belgium, we also needed a judge. In Belgium the judge is in control of the court hearing although his actual role is limited. The judge first asks the defendant if he understands the charges being made against him/her. The judge then allows the prosecutor to start with his pleading. Once the prosecutor is finished his presentation, the defense attorney can launch their pleading. At the end of the hearing, the judge sets the time frame for his final decision (normally four weeks after the hearing) and he could potentially ask some questions to the defendant. Our set up with the VR headsets does not allow for the evaluators to interrogate the defendant. It is, however, rather common in Belgian criminal proceedings for the judge not to ask questions. For the sake of realism in our videos, we needed an actual judge to control the flow of proceedings. Therefore, we selected a retired judge from the region of Hasselt to be in control of the cases like he would normally do. He was in fact sitting behind the 360 camera and participants could only hear his voice. This way the setting remained realistic and participants still viewed the cases from the perspective of the judge.

1.4.Selection of defendants

We needed three white defendants (each would participate in two cases) and three non-white defendants (each would participate in two cases) for our six cases. We decided to select actors that lived outside of the region where the evaluators are likely to reside (in order to avoid to be actors being recognized). We gave the actor-defendants wardrobe instructions so that the clothing of the pairs of actors was similar.

1.5.Selection of equipment and the shooting

We made use of a static 360 camera (OZO) to shoot the cases. The camera was positioned before the bench of the judge and was on the same height as the judge. This way the participants in the experiment observed the court room from the exact same angle as the actual judge. We hired a professional cameraman and a sound technician to make sure that the technical quality was optimal. Given that we shot the cases with a 360 camera, only the actors could be in the actual court room. The technicians were set up in a nearby courtroom where they could watch the camera footage on a monitor.

For each video we shot at least three full takes to make sure that there was a high quality take. After the video-shooting of a given case was completed, the video of the second defendant was shot in the empty courtroom. The production team replaced the former actor-defendant (in the first video) with the latter actor-defendant to generate identical VR videos in which only the defendants differed.

2. Organization of the experiment

2.1.Description of the equipment

We made use of 25 similar computers with powerful graphic cards. We tested various virtual reality headsets and decided to use the latest Oculus Rift headsets for the participants to wear during the experiment.

2.2.Description of the students

For the experiment we involved 156 students from the faculties of Economics and Law of Hasselt University. The Economics students were a mixture of bachelor and masters students who were enrolled in the Policy Evaluation course. The Economics group consisted of in total 89 students. They were randomly subdivided into four groups (given that we only had 25 headsets) to take the experiment. The economics students all took the experiment on the same day (November 21, 2017) and we made sure there was no interaction between the four groups of students. To avoid interaction between the groups, we put them in separate rooms and let them watch a movie (or follow a different course) while one of the other groups was taking the experiment. They were also instructed not to inform each other via mail or phone.

The law students are master students enrolled in Hasselt. They participated in the experiment as part of a course on Research Methods. There were 67 Masters' level law students enrolled in this course and they all took the experiment on November 22, 2017. They were randomly divided into three groups and again we made sure they could not interact with the other groups as long as the experiment was ongoing for one of the groups.

Participants were told that they would participate in an experiment which was an exercise in the adjudication and sentencing. We organized a random lottery (with 20 movie tickets) for students that participated in both the experiment and the follow up survey.

2.3.Description of the lawyer participants

In early July 2018 we randomly contacted lawyers via email and phone. We obtained lawyers' contact information from the website of the Limburg bar Association and only selected the ones listed under the category 'criminal law'. We contacted a little over 250 lawyers of which 36 decided to participate in the experiment in July 2018. The lawyers received the same information as the student participants and the flow of the proceedings was identical. The only difference is that the lawyers took the experiment individually during a 10-day period and they were instructed not to talk to anyone about the experiment.

2.4.Description of the case folders and the sequence of the cases

For each of the cases we made a case folder that contained the relevant information. We used a summary of the actual case folders from the cases, and changed the actual names, dates and places to make sure that no one could identify the actual defendant and victims. We also changed the ages of the defendants on the case folders so that they would match our actor-defendants. We produced six different cases (three burglaries and three assault) with two versions of each case (one with a white defendant and one with a non-white defendant). This means that in total we had 12 videos for our experiment and each defendant acted in two videos. We made six sets with different sequences of the videos and combinations in order to make sure that each participant would only see each defendant once.

2.5.Description of the anonymity and the website

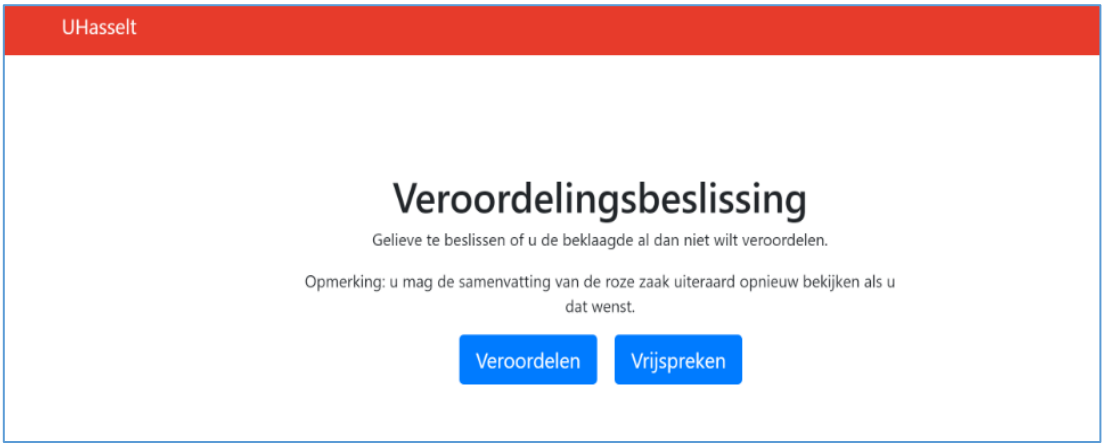
The participants are granted anonymity during the experiment and in the follow-up survey (see Section 3 of the Appendix). Upon entering the computer lab, students had to randomly pick a three digit number from a bowl. After picking the number, the students could randomly pick one of the 25 computers in the lab. Before they could start with the experiment they had to type in the three digit number in a custom made online tool.

For the purpose of the experiment we developed a website that would guide the participants through the experiment. The website contained clear instructions for participants on what to do (when to put on the headsets, when to take of the headsets, when to read a case folder, etc.). The website also communicated with the headset so that the videos would play automatically in the correct order. The experiment took 2 hours, with a five minute break after watching three videos.

Participants first had to read the case folder of a certain case. The folders were color coded to avoid confusion. After reading the case folder students could click on a button to start the video. They had 30 seconds to put on the headsets. After watching the video they could take off the headset and click on the button to go to the deliberation phase.

2.6.Description of assignment of sentences

After watching each case video, the students were faced with the choice to either acquit or convict the defendant. The picture below shows the screen students were shown during the deliberation phase.

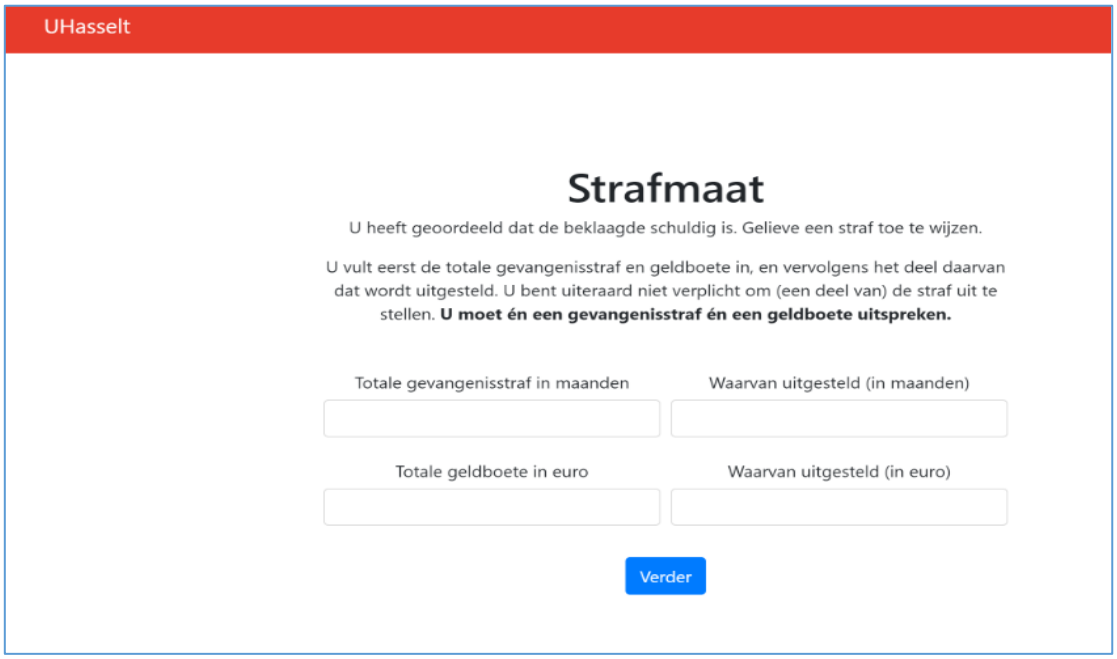


(*Translation: Conviction decision: Please indicate whether you want to convict or acquit the defendant. Note: You can read the case folder again if you want. Buttons: Convict; Acquit)*

If the evaluators acquitted the defendant, they would automatically proceed to the next case folder and they were instructed to read the next case folder. If they convicted the defendant they would go to the next screen to assign the sentences (which could be a prison sentence, a fine or a combination of both, depending on the circumstances of the case). For each case the evaluators could assign a sentence that was in accordance with the range prescribed by the penal code (for instance, a prison sentence between 12 months and 60 months). The Belgian criminal code also allows for the suspension of sentences (both for prison sentence and for fine). So the evaluators also had the option to assign a certain sentence and then suspend (part of) the sentence. For example, they could assign a prison sentence of 12 months and suspend 4 months of this sentence. This would mean that the defendant would go to prison for only 8 months, and the other 4 months were suspended. Similar to the actual judge, the participants had the choice of suspending (part of) the prison sentence and/or the fine. Both the law and econ students were given a small lecture (by the same lecturer) on the assignment of sentences before they participated in the experiment. If the penal code prescribes that the actual judge in a certain type of crime has to assign both a prison sentence and a fine or either one of them, we made sure the participants were faced with the same option. See the picture below.

We also made sure the students could not type in values for the postponed sentences that were higher than the total sentences. We also made sure that students could not go outside of the range prescribed by the penal code. The information on the sentence range allowed by the penal code was on the case folders (see the case folders at the end of the annex for the ranges of the sentences for each case).

After making the decision on acquittal or conviction and assigning the sentences for each of the six cases, the students were presented an overview of all their conviction decisions and sentences. They were than informed that they could make changes to their answers. We registered both the initial answers of the students as well as their changed answers.



(*Translation of text: You have decided that the defendant is guilty. Please assign a sentence. You first have to fill in the total prison sentence and fine, and next you fill in the part of the sentence that is suspended. You can of course not suspend (part of) the sentence. You have to fill in a prison sentence and a fine. Translation of boxes: Total prison sentence in months; postponed prison sentence (in months); Total fine in euro; postponed fine (in euro))*

(Translation: Overview conviction decision; # Case; Decision ; Prison sentence; Postponed prison sentence; Fine; Postponed fine)

2.7.Description of the testing

Two weeks prior to the experiment, we had an extensive testing phase. We tested the entire experiment with two groups of respectively 8 and 15 participants (who had no connections to the university). They tested both the functionality and the user friendliness of the online website and the headsets. This lead to the inclusion of a 1 minute instruction video on how to use the headsets in the beginning of the actual experiment. The participants also tested the clarity and terminology of the case folders which lead to minor changes in the wording.

3. The Follow-up survey

3.1.Drafting and testing the survey

We decided to organize a follow up survey for the participants. For the questions of the survey we made use of questions from the European Social Survey rounds 5 and 8. The survey contained groups of questions on the following topics: personal characteristics of the respondent, crime and migration, institutions, income and human values. We tested the survey on a group of 15 participants to make sure the questions were clear and the total duration of the survey took no more than 12 minutes. The survey was programmed in Qualtrics and was available via an online link. In the beginning of the survey the participants had to fill in the same three digit number in the experiment, this way we were able to match their answers and guarantee their anonymity.

3.2.Organization of the survey

We conducted the survey nine days to two weeks after the experiment during a lecture in the respective courses of the economics and law students. All students were informed to bring their laptops to fill in the survey. We send an email to students who were not in class with the link so that they could also fill in the survey. In total 156 of the 165 students that participated in the experiment filled in the survey. These 156 students constitute the sample used in the paper.
For the lawyer participants we registered the survey immediately after the experiment. Race questions were not included in the survey given to lawyers because all lawyers were white.

4. Testing the photos of defendants: descent

To make sure that the defendants that we used in the videos could be clearly identified as white or non-white, we organized a separate experiment with a group of 89 students from Hasselt University. These students were different than the ones who participated in the virtual reality experiment and they were from the first bachelor's year (freshmen). The students saw 7 pictures of males and had to indicate for each picture if they believed the person was of: (a) Western-European descent, (b) Middle Eastern or North African descent or (c) Asian descent. In the beginning of the experiment we explained to the students which type of countries we were thinking of. For Western-European we had Belgium, Netherlands, Germany and France in mind. For Middle-Eastern or North African we had Turkey, Morocco, Syria and Iraq in mind (they have the biggest communities in Belgium). For Asian descent we had China, Japan and South Korea in mind.

The pictures of our actor-defendants were projected on a large screen in the front an auditorium and each student had a separate voting box to indicate their responses. They were not allowed to discuss their answers with each other and had to respond within 30 seconds (all of them responded much faster). We added in a picture of a male that did not appear as a defendant in our videos.

The results from this photo experiment showed that the appearance of our defendants was very clear: there was almost no confusion as to whether our defendant stemmed from Western-European descent or Middle Eastern or North African descent. For three of our defendants all students correctly perceived their descent. For the other defendants either 86 or 87 of the 89 students gave the correct answer about the ancestry of the actor-defendants. This shows that there is little to no doubt as to the descent of our defendants.

5. Case descriptions

5.1.Burglary 1

Charge
Burglary, to the detriment of Frederik Smits and Bram Rogiers (victims). The following goods have been taken: a vehicle BMW 320cd, 100,00 EUR cash, a men's watch with a value of 50,00 EUR and spare keys of the BMW 320.

Defendant
- 20 years old
- Unmarried
- Unemployed

Criminal record of the defendant
- Refusal drug test and driving without driver's license: fine 2,400 EUR, 3 months driving ban.
- Driving without insurance and without driver's license, with vehicle that does not meet technical requirements of vehicles (motor vehicle inspection) and is not registered: fine of 2,400 EUR, 3 months driving ban.
- Burglary: 3-month prison sentence.
- Burglary by breaking, climbing in, or false keys: 8-month prison sentence.

Police report summary
- Victim Frederik Smits reported a burglary with breaking in his house.
- The forensics investigation for trace evidence in the house does not yield any results with the DNA and fingerprint databank.
- The burglar has gained access by ramming down a wooden door.
- 2 days after the burglary the victim informs the police that a friend has found the missing BMW.
- Forensics conducts an investigation of trace evidence on the car, and finds a fingerprint and DNA on exterior door.
- The fingerprint turns out to be the defendant's.
- House search (with consent defendant) at defendant's does not yield results (stolen goods not found).
- During interrogation the defendant denies having something to do with these facts, and does not know how his fingerprint ended up on the car. He does not know the victim.
- Victim says that he does know the defendant, however, the defendant has never been in his car with him.

Sentence range allowed by the penal code
- In case of conviction the defendant can be sentenced with a prison sentence from 1 month till 60 months and a fine between 156 EUR and 6,000 EUR.

5.2.Burglary 2

Charge
Burglary, to the detriment of John Peeters and Elke Deferm (victims). The following goods have been taken: a music installation of Bose, a golden watch from Ferrari, parfum of Mugler and Burburry, a HP laptop and 1,000 EUR cash.

Defendant
- 22 years old
- Unmarried
- Unemployed

Criminal record of the defendant
- Threatening to attack persons: 3-month prison sentence (suspended) and a fine of 600 EUR (suspended).
- Document fraud: 100 hours of community service.
- Indecent exposure: 2-month prison sentence (suspended).

Police report summary
- Victim John Peeters reported a burglary with breaking in his house.
- The forensics investigation includes taking trace evidence in the house and on an outside window. The DNA on the outside of the window matches with the DNA of the defendant.
- The defendant denies the facts and has no explanation why his DNA is on the window. He also claims to be physically not capable of breaking into the house due to a broken toe and a malfunctioning knee.
- The defendant admits to have been in the area of the break-in frequently at the time because he had a relationship with someone in the same area.

Sentence range allowed by the penal code
- In case of conviction the defendant can be sentenced with a prison sentence from 1 month till 60 months and a fine between 156 EUR and 6,000 EUR.

5.3. Burglary 3

Charge
Burglary, to the detriment of Tom Bamps (victim). The following goods have been taken: 2 HP laptops, 1 Apple tablet, 1 Rolex watch, 2 golden necklaces and 2,500 EUR cash.

Defendant
- 23 years old
- Unmarried
- Unemployed

Criminal record of the defendant

- No criminal record

Police report summary

- Victim Tom Bamps reported a burglary with breaking in his house. The victim came home from a night out and finds that the alarm system is not turned on. Upon entering the house he noticed that there had been a burglary. The neighbor indicates she noticed a black Volkwagen Passat standing in front of the house in the evening but she did not notice the number plate.
- The victim claims to have turned on the alarm system before he left the house. The alarm system appears to be working and there is no sign that the alarm system has been tampered with.
- The burglar(s) gained access through a small window in the back of the house that was not protected by the alarm system.
- The alarm system was only installed in the week prior to the break-in. The victim states that one of the technicians was asking a lot of questions on their social activities in the coming weeks. The victim saw him in their street a few days before the burglary.
- The forensic investigation turns up one finger print on the outside of the window that was used by the burglar(s) to gain access to the house.
- Both technicians that installed the alarm system are interrogated. One of them has an alibi for the night of the burglary. The second technician was fired from the company the day before the burglary under suspicion of stealing material from the company.
- The technician denies the allegations of breaking in. He states to have no financial problems and that his parents would support him if he had financial problems.
- The technician owns a black Volkswagen Passat (same model as seen in front of the house the night of the burglary). He claims not to have been in the street at the time of the facts.
- The fingerprint found on the window matches the fingerprint of the technician. He claims that the fingerprint was there because he installed the alarm system the week before the burglary.

Sentence range allowed by the penal code

- In case of conviction the defendant can be sentenced with a prison sentence from 1 month till 60 months and a fine between 208 EUR and 8,000 EUR.

5.4 Assault 1

Charge
Intentional assault of partner.

Defendant
- 25 years old
- Unmarried
- Unemployed

Criminal record of the defendant
- No criminal record

Police report summary
- Victim and defendant have a relationship for five years and have a 3-yeor old son together.
- There was an argument between the victim and the defendant, because he was not allowed to take the son to a family gathering.
- The victim tried to film the argument with her smartphone and the defendant grabbed the smartphone and smashed it. The victim states that the defendant than grabbed her and threw her on the table and on the ground. The defendant says she lost her balance and fell but that he was not behaving violently.
- The victim says she managed to get upstairs and call the police with another cell phone.
- The police see no apparent injuries on the victim. The police also see no signs on the table of a fight.
- The victim goes to the hospital on the same day and the medical record shows that she has pain in the neck, nausea and a tingling feeling in both her arms. The defendant claims that she had these complaints for an entire year and it has nothing to do with the argument.
- The victim claims that he has been aggressive before and hit her on multiple occasions in front of their son. The police was called to the house a year ago, but no police report was filed.
- The victim and the defendant decide to live in the same house for financial reasons. Once they sell the house, they will each go their own way.

Sentence range allowed by the penal code
In case of conviction the defendant can be sentenced with a prison sentence from 4 months till 24 months and a fine between 300 EUR and 1,200 EUR.

5.5. Assault 2

Charge
Intentional assault of partner.

Defendant
- 24 years old
- Married
- Employed

Criminal record of the defendant
- No criminal record

Police report summary
- Victim calls the police and claims that the husband has locked her in the house.
- Police enter the house through a window.
- Victim says that she and her husband have been problems for 6 months.
- Victim claims that her husband destroyed her clothes three weeks ago and strangled her in front of their 2 year old son. She shows the police pictures of her injured neck but she did not file it.
- Victim claims that her husband hit her and locked her in the house.
- The defendant claims that he did not hit his wife, and that she has a key so that she could have left the house. He claims to have never hit his wife and that she destroyed her own clothes three weeks ago.
- The defendants sister is interrogated and claims that the victim lived with her for a while because of the problems in her marriage. The sister claims that she could not believe that her own brother would use violence and states that the victim was aggressive herself.

Sentence range allowed by the penal code
In case of conviction the defendant can be sentenced with a prison sentence from 3 months till 24 months and a fine between 208 EUR and 1,600 EUR.

5.6 Assault 3

Charge
Intentional assault of partner.

Defendant
- 23 years old
- Unmarried
- Employed

Criminal record of the defendant
- Driving under influence: loss of driver's license for a month and 1,100 EUR fine.

- Driving under influence: loss of driver's license for 4 months.
- Document fraud: 7 months prison sentence (suspended) and fine (962.5 EUR).
- Online fraud: 6 months prison sentence and fine (600 EUR).

Police report summary

- Police are called to the house for a domestic dispute. The police have been at the house before for domestic disturbances while both partners are drunk.
- Police find a broken glass on the kitchen sink and hairs of the victim on the stairs.
- The victim goes to the doctor to document the injuries and provides the police with previous reports of injuries which she claims have been caused by the defendant.
- The argument started the previous day during a night out at a bar, when the defendant claimed that the victim was hanging around other men. The defendant left irritated and his partner stayed there. She did not come home that night, but only around noon the next day. The victim claimed to spent the night at a friend's place, while the defendant claims that she spent the night with another man. When she arrived at home, the couple started to have the argument.
- The victim claims that the defendant pulled her by her hairs throughout the house. After that she claims he pushed her down the stairs. The defendant denies to have hit her. He claims that she fell from the stairs herself (without him pushing her) and that he did aggressively tried to help her up by pulling her hair.
- Both of them admit to drink too much. The defendant claims that the victim has a real problem.
- The DA's office has tried mediation in this case to avoid it coming to court, but the mediation was not successful.
- The couple is back together at the moment of the trial.

Sentence range allowed by the penal code

In case of conviction the defendant can be sentenced with a prison sentence from 1 month till 12 months and a fine between 156 EUR and 600 EUR.

Appendix 2

Table A2-1

In-group Bias in Prison Sentencing- Models with Evaluator Fixed Effects

A: Regressions using trimmed sample			
	(1)	(2)	(3)
Minority Defendant	1.436***	1.417***	1.502***
	(0.456)	(0.442)	(0.446)
Minority Defendant	-3.848***	-3.544***	-3.626***
X Minority Evaluator	(1.346)	(1.232)	(1.204)
Early Trial	-1.294***	-1.566***	-1.323***
	(0.403)	(0.354)	(0.410)
Observations	634	634	634
Number of Evaluators	155	155	155
Evaluator FE	Yes	Yes	Yes
Trial FE	No	No	Yes
B: Regressions using untrimmed sample			
	(1)	(2)	(3)
Minority Defendant	0.887**	0.949**	1.077***
	(0.413)	(0.410)	(0.407)
Minority Defendant	-3.679***	-3.482***	-3.522***
X Minority Evaluator	(1.164)	(1.115)	(1.080)
Early Trial	-1.115***	-1.430***	-1.237***
	(0.358)	(0.320)	(0.377)
Observations	687	687	687
Number of Evaluators	155	155	155
Evaluator FE	Yes	Yes	Yes
Trial FE	No	No	Yes

Models in column (1) include Evaluator Fixed Effects. Models in column (2) include Evaluator Fixed Effects and Assault Case. Models in column (3) include Evaluator Fixed Effects and Trial Fixed Effects. Robust standard are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Appendix 2
Table A2-2
Bias in Prison Sentencing- Models with Evaluator Fixed Effects

A: Regressions using trimmed sample			
	(1)	(2)	(3)
Minority Defendant	1.001**	1.016**	1.091**
	(0.440)	(0.423)	(0.429)
Early Trial	-1.298***	-1.571***	-1.330***
	(0.407)	(0.361)	(0.418)
Observations	634	634	634
Number of Evaluators	155	155	155
Evaluator FE	Yes	Yes	Yes
B: Regressions using untrimmed sample			
	(1)	(2)	(3)
Minority Defendant	0.439	0.526	0.649*
	(0.396)	(0.391)	(0.391)
Early Trial	-1.125***	-1.441***	-1.257***
	(0.363)	(0.327)	(0.384)
Observations	687	687	687
Number of Evaluators	155	155	155
Evaluator FE	Yes	Yes	Yes

Models in column (1) include Evaluator Fixed Effects. Models in column (2) include Evaluator Fixed Effects and Assault Case. Models in column (3) include Evaluator Fixed Effects and Trial Fixed Effects. Robust standard are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Appendix 2
Table A2-3
In-group Bias in Fines- Models with Evaluator Fixed Effects

A: Regressions using trimmed sample			
	(1)	(2)	(3)
Minority Defendant	1.059***	1.057***	1.042***
	(0.218)	(0.216)	(0.208)
Minority Defendant	-1.345**	-1.318**	-1.206**
X Minority Evaluator	(0.525)	(0.522)	(0.524)
Early Trial	1.099***	1.075***	0.719***
	(0.241)	(0.240)	(0.249)
Observations	634	634	634
Number of Evaluators	155	155	155
Evaluator FE	Yes	Yes	Yes
Trial FE	No	No	Yes
B: Regressions using untrimmed sample			
	(1)	(2)	(3)
Minority Defendant	0.648***	0.653***	0.626***
	(0.211)	(0.210)	(0.207)
Minority Defendant	-1.177**	-1.160**	-1.083*
X Minority Evaluator	(0.560)	(0.562)	(0.580)
Early Trial	1.021***	0.994***	0.693***
	(0.233)	(0.232)	(0.240)
Observations	687	687	687
Number of Evaluators	155	155	155
Evaluator FE	Yes	Yes	Yes
Trial FE	No	No	Yes

Models in column (1) include Evaluator Fixed Effects. Models in column (2) include Evaluator Fixed Effects and Assault Case. Models in column (3) include Evaluator Fixed Effects and Trial Fixed Effects. Robust standard are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1

Appendix 2
Table A2-4
Bias in Fines- Models with Evaluator Fixed Effects

A: Regressions using trimmed sample			
	(1)	(2)	(3)
Minority Defendant	0.906***	0.908***	0.906***
	(0.206)	(0.204)	(0.196)
Early Trial	1.098***	1.073***	0.717***
	(0.240)	(0.239)	(0.249)
Observations	634	634	634
Number of Evaluators	155	155	155
Evaluator FE	Yes	Yes	Yes
B: Regressions using untrimmed sample			
	(1)	(2)	(3)
Minority Defendant	0.504**	0.512**	0.495**
	(0.200)	(0.198)	(0.196)
Early Trial	1.018***	0.991***	0.687***
	(0.232)	(0.232)	(0.240)
Observations	687	687	687
Number of Evaluators	155	155	155
Evaluator FE	Yes	Yes	Yes

Models in column (1) include Evaluator Fixed Effects. Models in column (2) include Evaluator Fixed Effects and Assault Case. Models in column (3) include Evaluator Fixed Effects and Trial Fixed Effects. Robust standard are clustered at the evaluator level. *** p<0.01, ** p<0.05, * p<0.1