Judicial Learning Curves and Their Implications: Evidence from Asylum Hearings

By
Clare Suter

Submitted in partial fulfillment of the requirements for the degree of
BACHELOR OF ARTS IN PUBLIC POLICY
at THE UNIVERSITY OF CHICAGO

Faculty Advisor: Jens Ludwig
Preceptor: Allison Kolar

04/20/2020

Acknowledgements

First and foremost, I owe my family tremendously for their encouragement. I completed the bulk of my writing at home during quarantine and couldn’t have asked for a better environment or support system. It didn’t hurt that if I ever got particularly stressed, I could always compare my situation to that of my younger sister, who, as the proclaimed opposite of a morning person, is currently battling her way through 6:30 a.m. online classes.

I’m also indebted to my boyfriend, who, by now, probably regrets having academic interests that are very similar to mine. The number of times I bombarded him with requests to talk through code, or read one of my sections for clarity is shameful, and he somehow remained a good sport through it all.

Finally, thanks to Jens Ludwig, not just for gracious advice on my thesis, but for initially sparking my interest in the intersection of data and policy. Without question, participating in his classes and research projects have had a profound impact on my undergraduate education, as well as my longer-term goals to contribute to a more equitable society. For this I am extremely grateful.

Abstract

This research examines the relationship between tenure and judicial decision-making, using a national longitudinal database of asylum court decisions. First, I investigate how an additional year of tenure affects the degree to which cognitive biases play a role in adjudications. In an extension of the regression model introduced in Chen, Moskowitz, and Shue (2016), I estimate the shape of a judicial "learning curve" using evidence of a specific cognitive bias: the gambler's fallacy. By observing how evidence of this bias decreases as experience increases by yearly increments, I find that the curve exhibits diminishing marginal returns, and identify the tenth year of tenure as the last year in which a judge can expect to improve their resilience to the gambler's fallacy. Next, I explore the implications of the cognitive bias learning curve for racial and cultural discrimination, using this tenth year benchmark to dichotomize experience. I use 9/11 as a natural experiment to compare inexperienced and experienced judges in their discrimination against asylum applicants from Middle Eastern Arab countries. Using the synthetic control method, I find that following 9/11, inexperienced judges significantly reduced their asylum grant rate for these applicants, while experienced judges showed no such statistically significant reduction. These results suggest that higher susceptibility to cognitive biases among inexperienced judges is costly beyond the vague notion that many asylum decisions are unduly influenced by irrelevant factors. More tangibly and disturbingly, this inability to separate signal from noise corresponds to the discriminatory treatment of certain groups of asylum seekers.

Table of Contents

Introduction

Judges hold a tremendous amount of power and responsibility in our society. Their actions routinely determine the fate of other people in a manner unlike that in other professions. At the end of the day, however, judges are only human, and therefore vulnerable to errors in judgement and decision-making. Judicial errors are a particular matter of concern because of the ability judges have to materially and wrongfully impact people's lives, under the premise of serving justice, or upholding the law. Though judges can be presumed to perform the best they can with the information provided, error is ultimately unavoidable, as many judicial decisions involve complicated predictions, incomplete information, or are affected by information irrelevant to the case at hand. Past research has generated critical knowledge concerning the manifestation of systematic judicial errors in a variety of settings.[1]

It seems, however, that not all judges err equally. A small but growing body of literature supports the notion that as judges accrue experience in court, their decision-making skills improve, and they commit fewer errors.[2] This type of "learning" has been observed across several different court settings, suggesting that improvements occur in the presence of varying levels of decision difficulty and complexity. In the context of judicial renderings on asylum cases, I attempt, in this analysis, to more closely examine the marginal benefits of experience on decision-making, and leverage the resulting insights to further explore the tangible implications of this relationship.

The growth of behavioral science as an academic discipline has yielded the discovery and understanding of a multitude of "cognitive biases", defined as systematic errors in the decision-

[1] Guthrie, Rachlinski, and Wistrich, "Blinking on the Bench."

[2] Arnold, Dobbie, and Yang, "Racial Bias in Bail Decisions." , "Chen, Moskowitz, and Shue, "Decision Making Under the Gambler's Fallacy."

making process that impact one's choices and judgements. [3] When cognitive biases influence an individual, the resulting *behavior* is not necessarily flawed. Rather, irrelevant information (commonly referred to as noise in this paper) acquires significant weight in the decision-making process, and consequently, relevant information (commonly referred to as signal in this paper) exerts less weight. From a research perspective, it is incredibly hard to measure the influence of cognitive biases on a decision, because doing so often requires identifying a departure from rational thought processes, which is difficult to prove, especially in judicial settings.

Fortunately, a particular cognitive bias lends itself well to empirical identification: the gambler's fallacy. The gambler's fallacy is the underestimation of the likelihood of streaks in a series of independent events. In the immigration court context, an asylum judge influenced by the gambler's fallacy would be more likely to deny asylum in a given case if she had granted asylum in the previous case. Such behavior is driven by the erroneous belief that two cases meriting asylum are less likely to be sequential than one case which merits asylum, and one which does not. In 2016, Chen, Moskowitz, and Shue published research specifically on the effect of the gambler's fallacy on asylum adjudications.[4] Using a regression to detect autocorrelation among judge decisions, the authors produce robust evidence that the gambler's fallacy does exert significant influence on immigration judges. Bearing special relevance to the inquiry in this paper, they find that less experienced judges (those with fewer than eight years of tenure) are more likely to exhibit gambler's fallacy-type decision patterns than their more experienced counterparts.

[3] Tversky and Kahneman, "Judgment under Uncertainty."

[4] Chen, Moskowitz, and Shue, "Decision Making Under the Gambler's Fallacy."

While this finding is informative, I seek, in this paper, a much more detailed understanding of the relationship between tenure and cognitive biases. In particular, what is the "shape" of this learning curve? Is it linear? Does it exhibit diminishing marginal returns? How large is the discrepancy in decision quality between those with the most tenure and those with the least? If the curve does indicate diminishing marginal returns, at what point do judges see the cognitive benefits of experience stagnate? Finally and importantly, does such a cut-off point have implications for judicial behavior that can be empirically observed? By using evidence of the gambler's fallacy to construct a year-by-year learning curve, I am able to offer answers to these questions, and provide novel contributions to the greater discussion on experience and cognitive biases.

To estimate the learning curve, I largely emulate the empirical approach of Chen et al.'s aforementioned paper. Critically, however, I formulate my regression model so as to facilitate an analysis of the marginal change in gambler's fallacy-type behavior by individual year of judicial experience. This allows for a much more granular understanding of the role of experience in this phenomenon than the one provided in Chen et al. (2016).

Following the construction of the learning curve, I turn my attention to the tangible costs of cognitive bias asymmetry across experience levels. Does this asymmetry drive meaningful differences in judicial behavior with respect to observable and relevant outcomes? In research published in 2018, Arnold, Dobbie, and Yang, find that inexperienced bail judges display relatively more racial bias than their experienced peers.[5] Arnold et al. speculate that this occurs because less experienced judges are more likely to rely on stereotypes, and are thus worse at predicting the genuine misconduct risk of a defendant. Because the paper does not directly

[5] Arnold, Dobbie, and Yang, "Racial Bias in Bail Decisions."

measure cognitive biases, the authors come short of attributing the experience-based differences observed in racial bias to a greater susceptibility to cognitive biases among inexperienced judges. By invoking evidence of the gambler's fallacy, I make possible a more direct analysis of the link between cognitive biases and more concrete forms of bias, such as systematic discrimination against applicants by protected class. Therefore, unlike in Arnold et al. (2018), the added analysis of the gambler's fallacy enables the assessment of the susceptibility to cognitive biases due to inexperience as a mechanism for discriminatory behavior.

Given that inexperienced judges appear to have greater difficulty limiting their considerations in adjudication to those strictly pertinent to a case, it seems probable that the noisy component of these deliberations might frequently involve factors such as protected class membership of asylum applicants. One would therefore expect less experienced judges to be (irrationally) more sensitive to the racial, religious, or national identities of the applications over which they preside, ostensibly engaging in more discriminatory behavior than their experienced counterparts. To explore this hypothesis, I leverage 9/11, an event which resulted in the proliferation of racially and religiously charged stereotyping and discrimination, as a natural experiment.

In the last decade, research has come to suggest that extraneous events, ranging from college football games to thunderstorms, impose an undue influence on decisions in judicial contexts.[6] These findings render naïve the view that adjudications are based solely on the law, or, at the very least based solely on deliberate considerations. The theory therefore accompanying this second phase of analysis is that exogenous shocks such as 9/11 flood the immigration court system with "noise", or irrelevant information. In the sense that the 9/11 attacks were instantly

[6] Danziger, Levav, and Avnaim-Pesso, "Extraneous Factors in Judicial Decisions"; Chen, "This Morning's Breakfast, Last Night's Game: Detecting Extraneous Influences on Judging"; Philippe and Ouss, "'No Hatred or Malice, Fear or Affection.'"

memorialized as a tragedy of historic national precedence, they remained highly salient among the American people for many months, and were rarely far from top of mind, even in unrelated contexts.

The explosion of emotional rhetoric and fearmongering following 9/11 drove a heightened focus on national security concerns, which resulted in certain religious groups and foreign nationals becoming widely stereotyped and villainized.[7] If a judge were only to consider the relevant information, or "signal", of an asylum application, one would not expect to see much difference in their decision patterns before and after 9/11. This is because an attack perpetrated by a terrorist group from a certain region, however horrifying and devastating, has little bearing on the merits of an asylum claim filed by someone from the same general part of the world. If a judge, however well-meaning, were influenced by the noise, chaos, and emotional rhetoric surrounding 9/11, she may subconsciously alter her decision patterns in ways harmful to certain applicants. Indeed, Mensah and Opoku-Agyemang (2018) exploit quasi-random variations in the timing of smaller-scale terrorist attacks and immigration court hearings, finding that these attacks negatively affect the chances of asylum applicants from Middle Eastern, North African, and predominantly Muslim countries.[8] Critically, research to date has yet to quantify the consequences of 9/11 for asylum applicants likely to face ensuing discrimination, let alone evaluate heterogeneity in such effects among judges.

Because judges who exhibit more cognitive biases are assumed to be more susceptible to the influence of noise, theory dictates that inexperienced judges are more likely to see group-

[7] Sikorski, Schmuck, and Matthes, "'Muslims Are Not Terrorists': Islamic State Coverage, Journalistic Differentiation Between Terrorism and Islam, Fear Reactions, and Attitudes Toward Muslims"; Human Rights Watch, *"We Are Not the Enemy": Hate Crimes Against Arabs, Muslims, and Those Perceived to Be Arab Or Muslim After September 11*; Rabby and Rodgers, "Post 9-11 U.S. Muslim Labor Market Outcomes"; Foner, *Wounded City*; Powell, "Framing Islam."

[8] Mensah and Opoku-Agyemang, "Innocent Until Stereotyped Guilty? Terrorism and US Immigration Court Decisions."

specific decision patterns disrupted by an exogenous shock that presumably affects widespread perception of that group. After leveraging the learning curve to establish an appropriate cutoff by which to dichotomize experience, I use the synthetic control method to test this theory, and diagnose any variation by experience group in reaction to 9/11. By evaluating asylum grant rate patterns before and after the 9/11 attacks, I assess whether inexperienced judge behavior following the attacks is indicative of increased bias against applicants from Middle Eastern Arab countries, as compared to that of the experienced group. This exogenous shock analysis illustrates the implications of the judicial learning curve by demonstrating the extent to which the inability to sift out signal from noise might actually give rise to a pattern of discrimination, and thus deprive asylum seekers a chance at a better future not just randomly, but *systematically.*

Background

The Asylum Application Process

Refugee asylum adjudications concern an extremely high-stakes decision, which for many applicants, is the deciding factor between life and death. Per the Refugee Act of 1980, which was designed to align US federal policy with the United Nations protocol, a refugee is formally defined as such:

> "any person who is outside any country of such person's nationality or, in the case of a person having no nationality, is outside any country in which such person last habitually resided, and who is unable or unwilling to return to, and is unable or unwilling to avail himself or herself of the protection of that country because of persecution or a

> well-founded fear of persecution on account of race, religion, nationality, membership in a particular social group, or political opinion."[9]

The task of an asylum judge then, is to evaluate whether the applicant at hand has a valid claim to such fear of persecution which directly stems from of any one of the five denoted factors: race, religion, nationality, membership in a particular social group, or political opinion. In the sense that the decision hinges on the potential dangers faced by the applicant should she be deported back to her country of origin, which is an unobservable outcome at the time of the decision, the judge faces a difficult prediction problem. The complexity of this decision is especially relevant to the analysis of cognitive biases in judicial contexts. One major consensus from decades of behavioral science research dictates that the more convoluted or abstract a given decision, the more likely a decision-maker is to rely on "automatic" thinking techniques or heuristics, which introduce bias into their judgement.[10]

Currently, the asylum application process can be initiated either affirmatively or defensively. In an affirmative application, a foreign national identifies herself to the Department of Homeland Security by filing for asylum. Several weeks later, the applicant is interviewed by a trained asylum officer in one of eight regional USCIS (US Citizenship and Immigration Service) offices. The interview is non-adversarial, with the asylum officer assuming an inquisitorial role. The applicant provides information about their identity, background, and reasoning for their pursuance of asylum and fear of persecution in their country of origin.

Decisions by asylum officers are reviewed by a supervisory asylum officer within the regional office before being released to the asylum applicant approximately two weeks after the

[9] Refugee Act of 1980, PL 96-212, 94 Stat. 102 (1980).

[10] Kahneman, *Thinking, Fast and Slow*.

interview takes place. If the asylum officer deems the applicant's interview materials credible, the applicant is granted asylum, which beside preventing deportation to the country of origin, conveys the benefits of work authorization, Social Security, and a path to permanent residence, among others. If the asylum officer either does not believe that the case is deserving of an asylum grant, or does not feel confident in extending a grant due to a lack of satisfactory evidence or any other hesitation, the case is referred to the immigration courts, and the applicant is placed into removal proceedings.

A foreign national who initially filed for asylum affirmatively, but was referred to the immigration courts after an interview with USCIS, will continue her pursuit of asylum by filing for relief under asylum in removal proceedings. Importantly, not all applicants in removal hearings originally filed affirmatively through USCIS. In contrast with the affirmative applicants, defensive applicants were apprehended in some way by the Department of Homeland Security, and placed directly into removal proceedings due to their illegal citizen status. Defensive applicants also have the option of defending against their removal by applying for the same forms of relief available to affirmative applicants, as discussed below.

Whether a case is affirmative or defensive has significant implications for the outcome of the case, seeing as affirmative cases tend to be viewed in a more positive light by judges, and are more likely to receive grants of asylum. The split between affirmative and defensive applications has traditionally hovered at around 40% and 60% respectively, but in the last five years, affirmative cases have plummeted, comprising only about 10% of today's asylum cases.[11] It is possible that this change in composition of cases reflects an increase in raids conducted by Immigration and Customs Enforcement (ICE), which would drive up the rate of defensive cases,

[11] "Asylum Decisions."

or a decrease in trust of the American government by non-citizens, which would drive down the rate of affirmative filings.

Removal proceedings occur in immigration courts, and are administrative proceedings which determine a foreign national's removability from the United States. As the name implies, cases for which the government establishes "removability" can result in the subsequent deportation of the foreign national involved. However, even if the government determines that the defendant is residing in the United States illegally, and is therefore "removable", there are several forms of relief for which a foreign national can apply. If granted, many of these protections provide temporary or permanent relief from deportation. In addition to asylum, alternate forms of relief include, most notably, withholding of removal, and relief under the Convention against Torture (CAT).

The standard for granting withholding of removal to the proposed country is whether removal to the proposed country would "more likely than not" result in persecution on account of race, religion, nationality, membership in a particular social group, or political opinion. Eligibility for withholding of removal is more difficult to prove than for asylum, because withholding of removal requires showing that is more likely than not (greater than 50%) that the applicant would be subject to persecution. Meanwhile, asylum requires demonstrating a "well-founded fear" of persecution, which is defined as a 10% likelihood of persecution. Unlike asylum, withholding of removal does not provide relief for an eligible individual's family members, no matter whether they are in the US or another country. Those who receive withholding of removal relief are not eligible for lawful permanent residence or citizenship, but can apply for work authorization in the US.

Relief via the Convention against Torture (CAT), involves protection from deportation for individuals who specifically fear torture in their home country. In order to qualify, an applicant has to demonstrate a clear probability (greater than 50% chance) that she will be tortured either directly by, or with the acquiescence of, the government of her country of origin. Ultimately, the nature of claims made by applicants of all three cases (asylum, withholding of removal, and CAT, also known as WCAT) is very similar, and many applicants feel that they are likely eligible for all three forms of protection.

Because removal proceedings in immigration court are not considered criminal cases, defendants (called respondents in this context) are not guaranteed the right to a lawyer. This is troubling given that many respondents do not have the means to seek legal representation. Also troubling is the fact that a trained attorney who often argues for deportation, represents the government in each of these cases, despite the lack of guarantee of counsel to the respondent. In recognition of the widespread inability among respondents to acquire representation, non-profits across the country have established networks by which they connect pro bono lawyers with respondents in removal proceedings. Often, these representatives prioritize respondents filing or eligible for protection under asylum or a similar form of relief.

Due to pro-bono efforts, the proportion of asylum cases involving unrepresented respondents is much less than one might expect, given the lingual and financial barriers which might preclude the average respondent from obtaining counsel. In the fiscal year 2018, for example, 84.4% of decided asylum cases involved respondents represented by counsel.[12] Representation is immensely consequential for achieving a favorable outcome in immigration court. Without representation, many in removal proceedings are unaware of the option to apply

[12] "Asylum Decisions and Denials Jump in 2018."

for various forms of relief, and all but seal their own fate by failing to do so. Among respondents who were never detained, those who were represented by counsel were nearly five times as likely to obtain relief if they sought it than those not represented by counsel.[13]

Once a case is formally in removal proceedings, the first court occurrence is a Master Calendar Hearing. This is a preliminary hearing at which the foreign national pleads to the initial charges filed against her in the removal proceedings, and if she chooses, formally requests the type of relief from removal sought. Critically, in order to be eligible for many types of relief, including asylum, the respondent must plead *guilty,* and admit removability, on at least one of the grounds of their charges. It is also worth noting that a respondent can apply for more than just one form of relief. Therefore, many seeking asylum will also apply for withholding of removal or relief under the Convention Against Torture, which convey similar benefits to those of asylum.[14] Notably, the judge presiding over the Master Calendar Hearing is the same judge who will later preside over the Merits Hearing. The Master Calendar Hearing closely resembles and functions like an arraignment in criminal proceedings.

The next, and usually the most important, stage of the removal proceedings process is the Merits Hearing. This is typically a very formal and adversarial evidentiary hearing on the record. The trial attorneys representing the government act as prosecutors, and attempt to disprove the applicant's eligibility for asylum or other forms of protection for which the respondent filed. Witnesses are sworn in on both sides, and both sides have the opportunity for direct and cross-examination. It is typical for the judges themselves to be very involved in questioning the respondent. At the end of the hearing, the judge will usually issue her oral decision, and grant or

[13] Eagly and Shafer, "A National Study of Access to Counsel in Immigration Court."

[14] "FY 2016 Statistics Yearbook."

deny all or a subset of the respondent's applications for relief from removal. On rare occasions, the judge will not issue a decision at the hearing, and will instead issue a written decision several days later.

If either the asylum applicant, or the DHS objects to the decision of the judge, they can choose to appeal the decision to the Board of Immigration Appeals (BIA), which is the appellate body within the Executive Office for Immigration Review, staffed by administrative judges appointed by the US Attorney General. The BIA has national appellate jurisdiction, and is composed of fifteen judges who gather in one location at the Executive Office for Immigration Review's (EOIR) headquarters in Falls Church, Virginia. Rather than hearing the case again, the BIA simply reads the transcript of the Merits Hearing which already occurred, and evaluates whether the immigration judge is likely to have erred. Parties cannot present additional evidence at this stage, but can advance specific legal arguments as to why the lower decision was incorrect. If the BIA also decides against the respondent, the respondent has the final option to file a petition for review before the Court of Appeals in the circuit in which the case was originally tried. If that application for relief is ultimately denied, deportation is scheduled immediately.

A Brief History of the Immigration & Asylum System

As previously discussed, the United States' current asylum framework was only established as of the Refugee Act of 1980. Since then, a number of significant policies have been implemented which have impacted the asylum process. In 1996, Congress passed the Illegal Immigration Reform and Immigrant Responsibility Act (IIRIRA), which generally constrained access to protection under asylum, and made it easier to deport foreign nationals.[15] It drastically expanded the list of crimes which made an immigrant eligible for deportation, and was effective retroactively. Those convicted of crimes on this list, many of which were nonviolent, were ineligible to argue their case or plead for relief in a court setting, and were immediately scheduled for deportation. The inability to engage in removal proceedings also became applicable to foreign nationals apprehended within 100 miles of the border.

Changes were made to detention policies for those in removal proceedings, so that drastically more people were detained before and between hearings. Beyond the obvious, detention is especially detrimental in its tendency to make it extremely difficult to acquire legal counsel. Perhaps most notably, IRRIRA required that all individuals seeking asylum at ports of entry be detained. The bar for obtaining any form of legal status was also dramatically raised. The legislation has continued to prevent many from effectively applying for asylum, and hindered the ability to obtain asylum on the basis of sexual orientation. It also imposed a one-year filing deadline, mandating that in order to be granted asylum, an applicant must file within one-year of her arrival in the United States. Even today, the United States government provides no official notice of this one-year deadline to those entering the country wishing to seek asylum.

Until 2003, the Immigration and Naturalization Service was largely responsible for administering asylum processes in coordination with the immigration courts. However, following the 9/11 attacks, the agencies overseeing immigration were reorganized. The Department of Homeland Security (DHS) was founded, and now acts as the umbrella organization for three smaller agencies. These agencies, operating as subsidiaries of DHS, are Immigration and Customs Enforcement (ICE), which deals with enforcement and deportation, Citizenship and

[15] Fragomen, “The Illegal Immigration Reform and Immigrant Responsibility Act of 1996.”

Immigration Services (CIS), which deals with immigration service and application processing, and Customs and Border Patrol (CBP) which deals with border inspection.

After declaring that the September 11th attacks had exposed the "vulnerabilities of [the US] immigration system", US Attorney General John Ashcroft signed off on a program titled the "National Security Entry-Exit Registration System" in September, 2002. It would come to be colloquially referred to as the Muslim registry.[16] The system was aimed at registering "certain types" of non-citizens within the United States, as well as increasing the screening of travelers and immigrants for "certain" counties.[17] All males from 25 chosen countries who were at least 16 years of age had to register.[18] Though the legislation obviously did not explicitly admit any goals pertaining to religious profiling, all but one country on the registry list was Muslim-majority.

The registry mandated that this select group of non-citizens to be subject to fingerprinting, photographs, and even interrogation. Furthermore, those who qualified as "people of interest" were required to frequently check in with immigration officials, and were monitored as they left the country to confirm that temporary visitors did not overstay their legal welcome. As a result of the program, those found to have violated provisions of the registry were arrested and fined, while over 13,000 individuals were placed in removal proceedings. Recent estimates place the total number of men and boys monitored by the Muslim registry above 80,000.[19] The ACLU, among other organizations, charged that the program was discriminatory on the basis of

[16] Shora, "National Security Entry Exit Registration System (NSEERS) Future Issues."

[17] Cainkar, "Targeting Muslims, at Ashcroft's Discretion."

[18] Countries of interest listed by the National Security Entry-Exit Registration System: Afghanistan, Algeria, Bahrain, Bangladesh, Egypt, Eritrea, Indonesia, Iran, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, North Korea, Oman, Pakistan, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, United Arab Emirates, Yemen.

[19] Penn State Law Immigrants' Rights Clinic and Rights Working Group, *The NSEERS Effect: A Decade of Racial Profiling, Fear, and Secrecy*.

ethnicity and religion. Though designed to identify terrorists, as of December 1st, 2016, not a single terrorism conviction had resulted from the program.[20]

Around the same time, in 2002, John Ashcroft implemented a series of sweeping reforms to the procedures and expectations of the immigration courts. In response to a tremendous backlog of cases, these reforms were ostensibly intended to make the removal proceedings process more efficient. Though seemingly counterproductive to the efficiency initiative, Ashcroft reduced the number of board members at the Board of Immigration Appeals, firing, at his own discretion, only appointees from the Clinton administration. He also directed the BIA to primarily follow a "summary affirmance" procedure, in which a single board member (rather than a panel of three) could decide to uphold an immigration judge's decision, and would not even be required to issue a short written opinion. Additional structures were set in place that made the task of reversing an immigration judge's denial significantly more laborious than that of concurring with the denial, which imposed a dangerous incentive structure on an already overburdened appellate body. Furthermore, Ashcroft directed that the current appellate backlog be cleared within half a year, which was calculated to translate to a board member deciding a case every 15 minutes.[21]

In 2005, the REAL ID Act was passed, which imposed further burdens on asylum seekers. It required that asylum seekers supply documents to verify their claims, unless they are able to demonstrate to their adjudicator why they cannot present such documents. The REAL ID Act has made it fundamentally easier for an adjudicator to deny an application for relief based on lack of credibility by the asylum seeker. In addition to imposing the document requirements, the

[20] Goodman and Nixon, "Obama to Dismantle Visitor Registry Before Trump Can Revive It."

[21] Kenney and Schrag, *Asylum Denied*.

legislation also required that applicants specifically demonstrate that their protected characteristic (in reference to the 1980 Refugee Act) was "at least one central reason" behind the actions of their persecutor in their country of origin.

In the last decade, ICE has taken a much more active role in seeking out foreign nationals that are residing in the US illegally, and detention of those apprehended by ICE has increased. ICE agents have begun sweeping and raiding homes, conducting sting operations, and detaining families with children.[22] In 2018, the Trump administration announced a "zero-tolerance policy" toward those who illegally enter the United States, dictating that all border crossers be referred to the Department of Justice and prosecuted for illegal entry.[23] This resulted in the separation of families who had crossed the border together, as the parents were placed in jail, while the children automatically become unaccompanied minors placed in custody or foster care. Public outrage about the inhumane nature of this policy ensued, however, and Trump eventually issued an executive order mandating that families be detained together.

Features of the Immigration Courts

The immigration courts are distinct from other judicial bodies in the United States in a variety of ways. First, the courts are characterized by an enormous backlog, which currently stands at over half a million cases waiting to be decided by approximately 450 immigration judges, all of whom are appointed by the US Attorney General.[24] Evidently, judges are under significant pressure to relieve the severity of the backlog, which is alleviated by rapid adjudication practices, often to the detriment of the asylum applicant. Additionally, there is very

[22] "ICEwatch: ICE Raids Tactics Map."

[23] "How The Trump Administration's 'Zero Tolerance' Policy Changed The Immigration Debate."

[24] Harris, "The One-Year Bar to Asylum in the Age of the Immigration Court Backlog."

little oversight applied to the judges' decision to grant or deny a given respondent's application for asylum. Judges are not required to issue a written decision, and typically announce their decision orally at the conclusion of the hearing. Though there is a system in place through which the public can file complaints against judges, judges almost never face consequences as a result.[25] Thus, judges are enabled to exercise significant discretion, which is especially impactful considering the typical scarcity of substantial evidence presented by asylees.

In contrast to other branches of law, asylum law hinges largely on factors that are incredibly difficult to analyze in a systematic or rigorous way. There is little application of technical details of the law, since the question at hand is almost always the same: does the respondent have a credible claim of persecution in their country of origin on the basis of any one of the five protected characteristics? There is usually little official material presented, and the decision hangs in the balance of the testimony given by the applicant, the receipt of which by the judge is inherently subjective. The heavy caseload and accompanying pressure to make progress with respect to the backlog, the lack of oversight, and the subjective nature of the hearing contribute to stunning levels of variation in decisions between judges.

In order to assess the merits of an asylum claim, a judge should understand the situation in the asylum applicant's country of origin to best decide whether her claim is both truthful, and warrants asylum under the 1980 guidelines. However, it is incredibly difficult to not only know, but keep track of civil unrest, hot conflict zones, and dangerous regimes throughout the world. Furthermore, immigration judges very rarely are staffed with law clerks, meaning that the busy judges themselves usually bear the burden of conducting such research.[26] Judges cannot

[25] Rosenberg, Levinson, and McNeill, "For U.S. Asylum Seekers, Some Judges Are a Better Bet than Others."

[26] National Association of Immigration Judges, Improving Efficiency and Ensuring Justice in the Immigration Court System.

reasonably be expected to know and use all of this information in an asylum decision, and understanding how a particular asylum applicant fits into the conflicts within her country of origin is an entirely additional problem. It is difficult to have much confidence in these decisions, because judges simply cannot possibly understand with depth all of the factors contributing to an asylum claim, especially given that they often make judgements based on a single testimony. To make things worse, research has indicated that asylum testimonies are unlikely to yield evidence that is helpful for adjudicating on the legal basis of asylum qualifications.[27]

Critically, immigration judges receive essentially no feedback, and have virtually no way of knowing whether they made the correct decision. If a judge choose to deny someone asylum, which results in the deportation of that person to their country of origin, she will never know whether that person ended up facing dangerous persecution that should have, in reality, qualified them for asylum. If a judge chooses to grant someone asylum, she will also never know what would have happened to that person in the case of deportation. The only way in which a judge *could* receive clear feedback on the accuracy of her decision might be if someone was publicly revealed to have lied in their testimony during their asylum hearing. Only recently, as revealed by a publication by Human Rights Watch in February 2020, have attempts been made to document what happens to those deported after having their asylum claim rejected. Though only focused on applicants from El Salvador, the report illustrates a grim truth, reporting the deaths of 138 former asylum seekers at minimum.[28] Additionally, a judge may regret her decision should she choose to grant someone asylum, and that person commits a crime on American soil, but it is critical to remember that asylum law is grounded in immediate humanitarian concerns.

[27] Keith and Holmes, “A Rare Examination of Typically Unobservable Factors in US Asylum Decisions.”

[28] “United States Deportation Policies Expose Salvadorans to Death and Abuse.”

Propensity to commit a crime does not make an asylum applicant any less likely to face persecution in their country of origin. Such predictions should therefore not change a judge's view on an asylum decision.

Literature Review

Cognitive Biases in Court

Much research has been conducted, especially in recent years, on the subject of departures from rational decision-making. A core facet of this literature concerns dual process theory, which posits that when confronted with a decision, the brain uses one of two distinct methods of reasoning. As defined by Kahneman (2011), the first style of reasoning, often dubbed the "automatic" type, can be understood as representing intuition, and reasons quickly and automatically, frequently drawing on emotional bonds in its reasoning process.[29] The second type of reasoning involves deliberation, and is much more controlled, slow, and conscious.

When automatic reasoning is involved in decision-making, miscalculations arising from cognitive biases are much more likely to occur, often resulting from the mental shortcuts used by this reasoning to arrive at an answer or decision. A substantial body of literature, including Thaler and Sunstein's "Nudge", explores the susceptibility of automatic reasoning to tweaks in the "choice architecture" of a decision environment, and demonstrates a myriad of circumstances in which people are driven to decisions that may be counter to their ultimate intention or objective function.[30] Decision environments with little structure, meaning that relevant

[29] Kahneman, *Thinking, Fast and Slow*.

[30] Thaler and Sunstein, *Nudge*.

information on which to deliberate is scarce, and decision-maker discretion is large, are known to facilitate the influence of cognitive errors significantly more than those with greater structure.[31]

Given their occupation as expert decision-makers, one might hope that judges do not fall prey to the cognitive biases so prevalent among laypeople. Evidence indicates that such hope is wishful thinking. In work by Guthrie, Rachlinski, and Wistrich, judges are subject to tests designed to expose vulnerability to five separate types of cognitive biases: anchoring, framing, hindsight bias, the representativeness heuristic, and egocentric biases (each bias is briefly defined in the footnotes for those unfamiliar with the concepts).[32] While the judges did appear to be somewhat less susceptible to framing effects and the representativeness heuristic than the average person (likely due to their occupational experience in making decisions), the tests determined that each one of the cognitive biases significantly impacted judicial decision-making by producing systematic errors in judgement.

Englich, Mussweiler, and Strack also examined evidence of the anchoring effect with respect to the judicial system, finding that irrelevant sentencing demands influenced the sentencing decisions of judges in criminal cases.[33] It should be noted that the research of both Guthrie et al. (2000) and Englich et al. (2016) was conducted in a lab setting. It therefore comes short of directly placing the influence of cognitive bias inside the courtroom. Documenting cognitive error in actual adjudications has proven a much rarer feat for researchers, but in recent years, a handful of ingenious study designs have found success in this endeavor.

[31] Kahneman, *Thinking, Fast and Slow*.

[32] Guthrie, Rachlinski, and Wistrich, “Inside the Judicial Mind.”; Anchoring: use of irrelevant information as a specific reference for evaluating a given decision. Framing: a dependence of preferences on the formulation of decision problems, such that a change to the wording of the question can drive a reversal in preferences. Hindsight bias: the tendency for people to perceive events which have already occurred as having been more predictable than they actually were prior to the events taking place. Representativeness heuristic: the similarity of objects or events confusing people’s thinking regarding the probability of an outcome. Egocentric biases: too heavy a reliance on one’s own perspective, and/or having a higher opinion of oneself than reality might require.

[33] Englich, Mussweiler, and Strack, “Playing Dice With Criminal Sentences.”

That cognitive shortcomings and biases pervade and disrupt our justice system is not necessarily a novel claim. As alluded above, several widely cited research initiatives have revealed spurious influences on adjudications. In one of the more shocking editions of these studies, proximity to a food break for parole judges accounted for a large portion of variation in the probability of a favorable outcome for the parolee.[34] This finding implies that minor hunger and associated grouchiness played a role in the judges' decisions, and thus the parolees' fates. Eren and Mocan (2018) uses 16 years of data to document the significant effect of the recent performance of a judge's alma mater, or local, football team on sentencing decisions.[35] In a pre-trial bail setting, Arnold, Dobbie, and Yang (2018) use instrumental variable techniques to detect evidence of racially-biased prediction errors among judges, which is argued to be distinct from taste-based discrimination, and instead reflective of genuine cognitive error, likely a result of stereotypes.[36] Pertinent to this paper is their finding of heterogeneity among experience levels: judges with less experience exhibit far more racial bias. This particular result supports the hypothesis that in my own research, inexperienced immigration judges will be found to exhibit more discriminatory behavior following 9/11 than their experienced colleagues.

Importantly, cognitive biases have specifically been demonstrated to play a role in asylum court settings. Many of these research efforts focus on the impact irrelevant events or factors have on judicial decisions. For example, Chen (2014) found asylum grant rates in US immigration courts to be influenced not only by the city's weather on the day of the decision, but also by the success of the court city's NFL team on the night before.[37] As discussed briefly in the

[34] Danziger, Levav, and Avnaim-Pesso, "Extraneous Factors in Judicial Decisions."

[35] Eren and Mocan, "Emotional Judges and Unlucky Juveniles."

[36] Arnold, Dobbie, and Yang, "Racial Bias in Bail Decisions."

[37] Chen, "This Morning's Breakfast, Last Night's Game: Detecting Extraneous Influences on Judging."

introduction, a 2018 paper by Mensah and Opoku-Agyemang assessed the degree to which racial profiling affect asylum decisions.[38] They evaluate how recent terrorist attacks in the US influence asylum grant rates for applicants from the Middle East, North Africa, and other predominantly Muslim countries, and are careful to control for a range of other factors which have been documented to affect asylum case outcomes, such as cloud cover, air pressure, and average due point. Their work is an indictment of judicial reliance on racial and religious stereotypes to make asylum decisions in the wake of these events.

As in the case of the pre-trial setting, Mensah and Opoku-Agyemang (2018) identified judgement errors that were found to be notably distinct from consistent, taste-based discrimination, due to their causal and temporal relationship with terror attacks. The judgement errors made following the terrorist attacks are consistent with a phenomenon called the availability heuristic. When people make decisions by relying on immediate examples or events that come to mind, they exhibit the availability heuristic.[39] The stereotypical association of certain nationalities and religions with recent terrorism events can, via the availability heuristic, serve to explain the drop in asylum grant-rates following attacks for applicants from Muslim-majority or Middle Eastern countries that is investigated in both Mensah and Opoku-Agyemang (2018) and my own research.[40] Such reactions are, of course, unfounded and statistically unsupported. Numerous studies have indicated that an influx of refugees into a population exerts somewhere between a null and positive effect on a range of crime types.[41] Ultimately, these existing papers focusing on the asylum setting are critical in developing the theoretical

[38] Mensah and Opoku-Agyemang, “Innocent Until Stereotyped Guilty? Terrorism and US Immigration Court Decisions.”

[39] Tversky and Kahneman, “Availability.”

[40] Keyes, “Beyond Saints and Sinners.”

[41] Chalfin, “What Is the Contribution of Mexican Immigration to U.S. Crime Rates? Evidence from Rainfall Shocks in Mexico”; Butcher and Piehl, “Cross-city Evidence on the Relationship between Immigration and Crime.”

framework of the exogenous shock analysis undertaken in this paper, but say nothing regarding the role of experience in these biases, which is my primary interest.

Though not specifically conducted in an asylum context, Phillippe and Ouss (2018) explore how television broadcasting of unrelated criminal justice events affects sentencing in France. Their findings indicate that professional experience and expertise mitigate the effect of irrelevant external information.[42] Importantly, these findings support one of the main theories driving this paper: that judges can improve decision quality over time by becoming less susceptible to errors arising from cognitive biases.

Bearing special relevance to my particular project is Chen et al. (2016), which indicates that gambler's fallacy plays a role in asylum adjudications. Judges were found to be more likely to grant asylum when they had denied it in the previous case, and vice versa.[43] Furthermore, experience mitigated the influence of the gambler's fallacy on judicial behavior, as judges with more tenure were less likely to demonstrate dependence between their subsequent decisions. However, their model defined a judge as experienced if the judge had amassed at least eight years of tenure. It appears that the use of eight years as a benchmark was driven by eight years being the median experience level. This choice is therefore largely arbitrary. I aim to recreate their analysis to allow for a much more granular and intentional understanding of experience's interactions with cognitive biases than has been produced by research to date.

Asylum Adjudications

The realm of immigration law that dictates asylum decisions has drawn notable attention from social scientist researchers over the years, which in no doubt has to do with its identity as

[42] Philippe and Ouss, "'No Hatred or Malice, Fear or Affection.'"

[43] Chen, Moskowitz, and Shue, "Decision Making Under the Gambler's Fallacy."

an incredibly high-stakes environment. A more thorough description than that provided in this paper of the asylum application process is given by Ramji-Nogales, Schoenholtz, and Shrag (2007), which also sheds light on the drastic variations in asylum rulings between and within immigration courts.[44] Their 2007 publication, which a similar report by the Government Accountability Office (GAO) quickly corroborated, provoked outrage and dissatisfaction with the asylum adjudication system, as it revealed gigantic disparities in grant rates, even when adjudicators in the same court considered large numbers of applications from the same country.[45] In response, the Executive Office for Immigration Review held training sessions to address the disparities, in addition to creating an apparatus though which the public could file complaints against immigration judges. In a 2016 GAO report, however, it was concluded that these efforts were fruitless – disparities in grant rates were still extraordinarily large.[46]

For example, it was calculated that Colombian asylum applicants whose cases were adjudicated in the federal immigration court in Miami had a 5% chance of prevailing with one of that court's judges, and an 88% chance of prevailing before another judge in the same building. Clearly, the asylum decision is accompanied by an enormous amount of discretion, the size of which would be presumed wildly incomprehensible for other high-stakes court contexts, such as those concerning death sentences. In 2008, after compiling information on variations in asylum outcomes, the US Commission on International Religious Freedom's Annual Report remarked that these outcomes appeared to depend largely on chance assignment to a particular immigration judge.[47] Thus, in a manner so opposite to that of the fair and idealized American justice system,

[44] Ramji-Nogales, Schoenholtz, and Schrag, "Refugee Roulette."

[45] U. S. Government Accountability Office, "U.S. Asylum System."

[46] U. S. Government Accountability Office, "Asylum."

[47] US COMMISSION ON INTERNATIONAL RELIGIOUS FREEDOM, 'Annual Report of the US Commission on International Religious Freedom' Keith and Holmes, "A Rare Examination of Typically Unobservable Factors in US Asylum Decisions."

the most critical moment in an applicant's case is not during the applicant's hearing or testimony, but rather in the applicant's random assignment to a judge, in which she has no power or say.

A 2009 study leveraging data from asylum hearings in Texas noted that most of the asylum seekers did not have substantial levels of evidence to present, which other legal analyses confirmed to be the norm in these cases.[48] The factors from this study that did appear to influence asylum decisions seemed to have little to nothing to do with the legal basis of asylum, and were often personal characteristics of the applicants, such as being female, or married. Ultimately, the consensus from research on asylum proceedings supports the notion that the decisions faced by the immigration judges are incredibly complex, rarely involve useful evidence or information, and allow for what seems like an unprecedented amount of judicial discretion.

9/11 & Anti-Muslim Sentiment

Few events have exerted as great an exogenous shock to the US immigration courts as did the September 11th terror attack in 2001. The attack was unprecedented in American history, and shocked the world. Its rippling effects were virtually inescapable, as there were broad implications for foreign policy, financial markets, and political races. It has firmly maintained a place in the American national consciousness, even almost two decades later, and there exists a wealth of literature dedicated to the ways in which it reshaped public rhetoric and perspective. The short-term reaction by many groups, including the US government on many accounts, was to regard Muslims and Middle Easterners in the United States unanimously with suspicion and distrust.[49] Despite the far-reaching and clearly documented anti-Muslim sentiment associated with the aftermath of the attacks, my research is poised to be the first to analyze how 9/11

[48] Keith and Holmes. "A Rare Examination of Typically Unobservable Factors in US Asylum Decisions"

[49] Foner, *Wounded City*.

affected the success of asylum applications filed by those hailing from Middle Eastern Arab countries. Below I provide an overview of previous research on the September 11 attacks which motivate such an inquiry.

Unsurprisingly, the event was a focal topic in the media for months, and even years, and recent research inquiries have indicated that since 9/11, many American media outlets are guilty of reporting in a way that connects terrorism to Islam, thus creating a fear of the "other", and perpetuating Islamophobia.[50] This connection, or failure to explicitly make clear the vast distinctions between Muslim communities and Jihadist terrorist groups has proven extremely damaging to public opinion and sentiment regarding the Muslim population.[51] It is important to note that the government and other private sector organizations did take some initiative in combatting this discrimination. Typically they did so by leveraging media outlets to convey messages directly intended to stem the tide of hate, and correct ill-informed prejudice.[52] Generally though, these efforts did not appear to succeed in eclipsing the effects of the much more consistent, though often subconscious and subtle messaging, promoted by the mainstream media regarding the Muslim identity's inextricable links to Islamic terrorism.

Research revealed a significant spike nationwide in anti-Muslim hate crimes following 9/11, although this increase was notably insignificant in both New York City and Washington D.C., the primary targets of the attacks.[53] Reports conducted by multiple organizations concurred in estimating that in 2001, the occurrence of these hate crimes rose by a whopping 1700%.[54] The

[50] Powell, "Framing Islam."

[51] Sikorski, Schmuck, and Matthes, "'Muslims Are Not Terrorists': Islamic State Coverage, Journalistic Differentiation Between Terrorism and Islam, Fear Reactions, and Attitudes Toward Muslims."

[52] Arab American Institute, *Healing the Nation : The Arab American Experience after September 11*.

[53] Byers and Jones, "The Impact of the Terrorist Attacks of 9/11 on Anti-Islamic Hate Crime."

[54] Federal Bureau of Investigation, "Crime in the United States 2001"; Human Rights Watch, *"We Are Not the Enemy": Hate Crimes Against Arabs, Muslims, and Those Perceived to Be Arab Or Muslim After September 11*; Ibish, "Report on Hate Crimes and Discrimination against Arab Americans: The Post-9/11 Backlash, 9/11, 2001-October 11, 2002. Washington, DC."

effects of post-terrorism animosity toward Muslims have also been observed in less overtly criminal settings, as studies conducted in the US, as well as the UK, examined the labor market consequences of these attacks.[55] Rabby and Rodgers (2011) found that particularly for younger Muslim men, employment rates decreased significantly in the years following 9/11, as well as after a 2005 bombing in London.[56] Such findings point to widespread, anti-Muslim discriminatory reactions to attacks perpetrated by jihadists, which appear to be far from contained within the epicenters of the physical attacks.

The xenophobic response to those perceived to be "foreign", or "non-authentically" American was not an anomalous event in American history. Even to groups in need, or experiencing persecution, the American public has historically demonstrated a collective, inhumane apathy to groups with whom they do not personally identify. In 1938, for example, though nearly all Americans condemned the Nazi regime's crimes against the German Jewish population, the very week after Kristallnacht (The Night of Broken Glass/the November Pogrom), a Gallup poll reported that 72% of American respondents answered "No" to the question: "Should we allow a larger number of Jewish exiles from Germany to come to the United States to live?"[57] The significant role of xenophobia and apathy in contexts that ideally should be dominated by humanitarian sentiments is therefore far from unprecedented. Thus, though many asylum seekers from Muslim countries following 9/11 were fleeing the regimes of the very terrorist organizations feared by Americans, American historical tendencies lend themselves to the attitude that the plight of these groups is easy to ignore, especially in return for a perception of improved national security.

[55] Rabby and Rodgers, "Post 9-11 U.S. Muslim Labor Market Outcomes."

[56] Rabby and Rodgers, "The Impact of 9/11 and the London Bombings on the Employment and Earnings of U.K. Muslims."

[57] Gallup Inc, "American Public Opinion and the Holocaust."

Research by Skitka et al. (2006) also suggests that the 9/11 attacks drove Americans to identify as more politically conservative, likely due to the emphasis by the Republican party of national security interests through the promotion of a "War on Terror"[58] Such policies were largely viewed as addressing the fear and unrest felt by many Americans after the attacks.[59] If these sentiments were internalized or echoed by judges, it may have led to an increased prioritization of "national security concerns" in the admission of asylum, which would theoretically lead to a decrease in the amount of asylum applications approved, due to fear or suspicion of malintent. As discussed, terrorist attacks have been shown to result in decreased asylum grant rates for Muslims and Middle Easterners in America due to the use of stereotyping by judges.[60] Critically, this research found that the effects of terror attacks were confined to the immigration courts, and were not observed in other judicial settings, implying that by virtue of their interactions with foreign nationals, immigration hearings are widely perceived to operate in a context of elevated relevance to national security concerns. This is somewhat ironic given that the overwhelming majority of terrorist attacks carried out on American territory have been perpetrated by US-born citizens.[61]

Attempts to quantify the impact of 9/11 have been complicated by the challenge of finding appropriate comparison groups against which to measure the effect on a given outcome. That is, 9/11 pervaded essentially every corner of society, meaning that a counterfactual in which the attacks did not happen proves virtually impossible to construct. Despite these challenges, a number of studies leveraged creative research designs in order to determine its effects on a wide

[58] Skitka et al., "Confrontational and Preventative Policy Responses to Terrorism: Anger Wants a Fight and Fear Wants 'Them' to Go Away."

[59] Bonanno and Jost, "Conservative Shift Among High-Exposure Survivors of the September 11th Terrorist Attacks."

[60] Mensah and Opoku-Agyemang, "Innocent Until Stereotyped Guilty? Terrorism and US Immigration Court Decisions."

[61] Bergen et al., "Who Are the Terrorists?"

range of outcomes. In one such study, Lee et al. (2005) estimate the impact of the attacks on US air transport passenger demand using a technique titled Single Interrupted Time Series Analysis.[62] In this method, the temporal trends and fluctuations of passenger demand before the attacks were fit to an autoregressive time series model, which was projected into the post-period to serve as a reasonable counterfactual, or estimate of demand, in an alternate reality in which the attacks did not occur. In research attempting to assess the impact of 9/11 on tourism to the US, Gut and Jarrell (2007) employ a general autoregressive distributive lag model (ADLM) in order to test for different data generating processes after the attacks.[63] The Single Interrupted Time Series and ADLM techniques are somewhat unideal, however, in that they do not use control groups and can involve significant extrapolation.[64]

The previously mentioned work of Rabby and Rodgers (2011) regarding labor market consequences for Muslim men in the United States and United Kingdom adopted a different framework, which utilized a Difference-in-Difference design.[65] In these models, the treated condition was being Muslim in the period following the terrorist attacks, while the control group in both the pre and post period, was all other ethnicities. By narrowing the focus of interest to one specific cohort (Muslim men), for which theories of discrimination predict a change in outcome, the comparison group can be made up of the remaining ethnic groups, who can be more safely presumed not to be perceived differently in accordance with terror attacks, due to a lack of religious, ethnic, or cultural stigma associated with the attacks. I will be leveraging a similar framework in my analysis of post 9/11 anti-Muslim discrimination.

[62] Lee, Oh, and O'Leary, "Estimating the Impact of the September 11 Terrorist Attacks on the US Air Transport Passenger Demand Using Intervention Analysis."

[63] Gut and Jarrell, "Silver Lining on a Dark Cloud: The Impact of 9/11 on a Regional Tourist Destination."

[64] Baicker and Svoronos, "Testing the Validity of the Single Interrupted Time Series Design."

[65] Rabby and Rodgers, "The Impact of 9/11 and the London Bombings on the Employment and Earnings of U.K. Muslims"; Rabby and Rodgers, "Post 9-11 U.S. Muslim Labor Market Outcomes."

Identification Problems & Solutions: Evaluating Judicial Performance

The complexity of the decision coupled with the significant amount of discretion afforded to judges renders asylum adjudications a context in which one would theoretically expect cognitive biases to wield significant influence. A fundamental problem of any research evaluating judicial behavior, however, is the difficulty associated with decisively labelling any given decision as correct or incorrect. There is no ground truth against which to compare empirical judicial decisions. This challenge stems largely from unobservable counterfactuals, which concern the outcomes associated with the judge making an alternative decision.

For example, it is impossible to know what would have happened had an asylum seeker who was denied asylum and deported been granted asylum, or vice versa. Importantly, in asylum cases, it is also essentially impossible to know the ultimate outcome of the decision that was actually made. If, for instance, an asylum seeker were deported back to her country of origin and ended up facing harm from the persecution she claimed to fear in her asylum application, it is clear that the judge made an incredibly costly error in not granting that applicant asylum. However, there exist virtually no channels of information by which a judge (or researcher) would learn of this mistake.

Additionally, it is difficult to rule out the possibility that differences in case outcomes, for which observed covariates in the data are identical, are due to the judge's observation of private information, which provides valuable signal regarding the merits of a case.[66] In other words, two cases may look identical on paper but play out very differently in the courtroom, provoking the judge, who observes these practical differences, to issue different decisions. If the data only contains information on a limited number of the true variables relevant to underlying case

[66] Kleinberg et al., "Human Decisions and Machine Predictions."

quality, how can one assert that a given decision was mistaken or irrational, rather than the product of careful observations made by the judge with respect to a factor not captured by the data? Several econometric strategies have attempted to dodge this problem by relying on statistical assumptions stemming from quasi-random assignment of cases to judges. Additionally, researchers have increasingly relied on a distribution of judicial outcomes under the null hypothesis of decision-making in the absence of cognitive or other biases. A variety of creative techniques have ultimately made possible the analysis of judicial performance and behavior in a range of settings.

In scenarios in which judicial decisions are frequently appealed, and are occasionally overturned, Norris (2018) proposes a measure of judicial performance dubbed "consistency"[67] This metric tracks the portion of a judge's decisions that are overturned, and ultimately uses this disagreement between a judge and a higher court, or a board of appeals, as a proxy for accuracy. Norris illustrates the use of this measure on Canadian asylum data, which generally resembles the characteristics of US asylum data. While the consistency measure can prove useful for comparing judges, it is not as useful for my inquiry, which is more focused on the specific role of cognitive biases in performance.

A non-parametric approach leveraged by Abrams, Bertrand, and Mullainathan (2012) seeks to assess judicial differences in racial bias by measuring the between-judge variation in the difference in incarceration rates and sentence lengths between African American and white defendants.[68] A Monte Carlo simulation is then performed to construct the appropriate counterfactual, in which race does not affect judicial sentencing behavior. This simulated

[67] Norris, "Judicial Errors."

[68] Abrams, Bertrand, and Mullainathan, "Do Judges Vary in Their Treatment of Race Conference."

behavior is then compared to the empirically observed judicial behavior. Because the empirically observed behavior does not lie within the distribution of simulated behavior, it is concluded that race affects incarceration outcomes. This approach is advantageous in that it controls for unobservable variables, and does not require a model to be correctly specified. I find the technique of simulation under the null hypothesis to be extremely helpful in determining the significance of my findings with respect to grant rate behavior following 9/11.

In general, the random assignment of asylum applicants to judges is a statistically virtuous feature of studies involving observational court data, and has made possible countless research efforts focused on the judicial system. It allows for the assumption that each judge receives a similar distribution of cases in terms of ground truth merit. Statistical differences between judges are therefore more easily attributed to differences in decision behavior rather than to differences in applicant or defendant behavior. In the gambler's fallacy-motivated research of Chen et al. (2016), a regression-based approach is leveraged using case-level data, in which the outcome is whether or not the judge granted asylum in the particular case.[5] If the gambler's fallacy did not bear on such a decision, one would expect that the most recent decision by that judge is not a statistically significant predictor in this equation. Their research demonstrates, however, that the lagged decision predictor *is* statistically significant, and robust to a variety of subsamples and specifications of the model. Importantly, Chen et al. (2016) calculate their estimates using clustered standard errors (by judge), which are necessary for data structures that involve multiple data points over time associated with the same individuals.[69] This accounts for the unobserved correlation between observations stemming from the same judge over time.

[69] Bertrand, Duflo, and Mullainathan, "How Much Should We Trust Differences-In-Differences Estimates?"; Abadie et al., "When Should You Adjust Standard Errors for Clustering?"

As discussed previously, Chen et al. (2016) test for the effect of experience on this cognitive bias, using a single experience dummy interacted with the lagged decision term. Their estimates of this term indicate that, given a previous grant, judges with less than eight years of tenure are, on average, between 3.3% and 4.6% less likely to issue a grant for their current decision than judges with eight years of tenure or greater.[70] This estimate is not specifically informative regarding the gap that might exist between the very oldest and very youngest judges. It does not provide insight into the gaps that might exist between each consecutive year of tenure. Though I will be emulating Chen et al. (2016)'s approach to identification, I aim to expand the conclusions that can be drawn surrounding experience's role in cognition by incorporating specifications that allow the experience variable to be observed in a non-dichotomous context.

Finally, a non-parametric technique which has proved promising, not in judicial contexts specifically, but certainly in criminal justice contexts more generally, is known as the synthetic control method.[71] This method is especially useful in research settings where there is one treatment unit, and several control units. It also adds significant value to identification problems in which the ideal control unit(s) is either not clear, or unsuitable as a control unit due to a failure of identification assumptions. As the name implies, the synthetic control method involves the construction of an entirely novel, or "synthetic", control unit from linear combinations of existing controls, so as to best approximate the behavior of the treatment unit before the policy or intervention of interest. The treatment unit, which has undergone treatment following the

[70] Chen, Moskowitz, and Shue, "Decision Making Under the Gambler's Fallacy."

[71] Rydberg et al., "A Quasi-Experimental Synthetic Control Evaluation of a Place-Based Police-Directed Patrol Intervention on Violent Crime"; Saunders et al., "A Synthetic Control Approach to Evaluating Place-Based Crime Interventions"; Pyun, "Exploring Causal Relationship between Major League Baseball Games and Crime"; Donohue, Aneja, and Weber, "Right-to-Carry Laws and Violent Crime: A Comprehensive Assessment Using Panel Data and a State-Level Synthetic Control Analysis."

intervention, is then compared to the synthetic control unit, which is assumed to behave as the treatment unit's counterfactual, or as the treatment unit in a world where the intervention of interest had not occurred.

Abadie, Diamond, and Hainmueller (2010) first popularized the technique in research estimating the effect of a cigarette tax in California.[72] Since then, it has been used to estimate the impact of terrorism on economic growth, of universal basic income programs on labor market participation, and of ethnically motivated civil war on economic wellbeing.[73] Furthermore, the synthetic control approach has been used in studies of heterogenous reactions to events, which is exactly my goal in assessing reactions along experience levels to 9/11. Because the method is non-parametric, standard error estimation, which has direct implications on the significance of estimates, is somewhat less straightforward than in regression contexts. Despite this, fairly conventional methods have developed for reporting significance which borrow from the logic of permutation inference.[74] I find the synthetic control method to be a useful strategy in my assessment of the difference in reactions between experienced and inexperienced judges to 9/11.

Learning Curve

In this section, I detail the data and methods used in, as well as the results from, my analysis of the relationship between the tenure of immigration judges and their exhibition of the gambler's fallacy. I perform this analysis using yearly increments in tenure to gain a nuanced understanding of marginal improvements in decision quality, and construct a judicial "learning

[72] Abadie, Diamond, and Hainmueller, "Synthetic Control Methods for Comparative Case Studies."

[73] Abadie, Alberto Abadie, and Gardeazabal, "The Economic Costs of Conflict: A Case Study of the Basque Country"; Rabby and Rodgers, "Post 9-11 U.S. Muslim Labor Market Outcomes"; Costalli, Moretti, and Pischedda, "The Economic Costs of Civil War: Synthetic Counterfactual Evidence and the Effects of Ethnic Fractionalization."

[74] Abadie, Diamond, and Hainmueller, "Synthetic Control Methods for Comparative Case Studies."

curve". This analysis is a precursor to that of the relationship between judicial experience and anti-Muslim discrimination following 9/11.

Data

Though my learning curve model largely follows the empirical specification of Chen et al. (2016), I was not able to use the original data from their paper, as it is unavailable for public use.[75] I proceed by using freely available asylum data, and attempt to replicate their model by recreating their covariates. Notably, I am unable to conduct this replication with 100% accuracy, due to a difference in data source and, consequently, in available variables. In general, however, I succeed in constructing all but one of their controls, and end up running my model on a sample size more than twice as large as that in Chen et al. (2016).[76]

I acquired my asylum data through a Freedom of Information Act public release file on the Executive Office of Immigration Review website.[77] It includes case, proceeding, and application-level data on all asylum adjudications conducted in the United States from January 1990 through early December of 2019. Variables in my data include the following: defendant nationality, a unique ID for each case, a unique ID for each court proceeding, a unique ID for each relief application filed, legal representation indicator, the time and date of the Merits Hearing, judge name and ID, the immigration court in which the case was heard, family size, and the outcome of the case, as well as specific outcomes for individual applications for relief.[78] The

[75] Chen, Moskowitz, and Shue, "Decision Making Under the Gambler's Fallacy."

[76] The model of Chen et al. (2016) with the largest sample size had 150,357 observations. My learning curve model has 345,083 observations.

[77] The data was downloaded from the EOIR website on December 11, 2019. Because the link is frequently updated, it is no longer possible to replicate the exact download of the data in this paper, but a more current version can be accessed at this link: www.justice.gov/eoir/frequently-requested-agency-records.

[78] The original data was comprised of several tables downloaded from a relational database. Each Case ID was associated with one or more Proceeding IDs, which was associated with one or more Application IDs. Judge decisions in time on relief applications can be thought of as happening at the proceeding level, which is why decisions were analyzed at this level of data. To flatten this data structure and obtain one decision per proceeding, I collected a list of each decision at the application level associated with a particular Proceeding ID. If any decisions within one of these lists was a "grant", I designated that proceeding to have resulted in a grant. If no grant codes were present, but a deny code

family size variable was not present in the original data – I manually constructed it using a data table that linked the lead family member in each case to their dependents.

I merged this data with hand-collected biographical information on judge start dates. This allowed me to create a moving judicial experience variable by subtracting a judge's start date from the given hearing date.[79] Because assessing presence of the gambler's fallacy involves evaluating the degree to which a previous grant decision affects a judge's decision to grant or deny her currently adjudicated application, the outcome variable in the gambler's fallacy analysis will be the immigration judge's decision on a given asylum application.

For the purpose of this analysis, withholding from removal, and protection under the convention against torture (WCAT) are both considered a positive outcome, and therefore a "grant". Many refugees in removal proceedings apply for withholding from removal and WCAT in addition to asylum; in fact, it is uncommon to file only for asylum.[80] As discussed previously, these additional forms of relief provide similar benefits to those of asylum. It is therefore sensible to categorize such decisions as grants rather than a denials, because a grant of WCAT, for example, implies a categorically different and more positive outcome for an applicant than a denial of all relief applications, and immediate deportation.

In order to correctly test for evidence of the gambler's fallacy, it is critical that the data used is comprised of accurately ordered observations, and only includes decisions that were directly adjacent to one another in a given judge's schedule. This is because in order to accurately estimate the effect of a judge's prior decision on her current decision, I must confirm

was present, I designated that proceeding to have resulted in a "deny". If no known grant or deny codes were present at the application level for a given proceeding, it was designated as an "other", and eventually discarded from the final data used for the analysis.

[79] Biographical information, including start dates, could only be obtained for 452 of the 879 judges in the original data. The data was subsequently restricted to decisions made by judges with known-start dates.

[80] "FY 2016 Statistics Yearbook."; The applications for all three forms of relief are convenient to file simultaneously – they are all on the same document: USCIS Form I-589

that the prior decision in the data is indeed the decision that immediately preceded the next. Otherwise, the true previous case's effect will be completely misattributed. I therefore restrict my data to cases that can be ordered on a known timeline, discarding observations which are associated with a specific date, but not a specific time.

While the majority of relief applications had recorded decisions of either "deny" or "grant", about a third of all decisions were labelled "other", which signaled that a variety of possible outcomes such as change of venue, failure to appear, or withdrawal of applications had transpired. Because it was impossible to determine the outcomes of interest in these cases, they were excluded from the data, as were observations immediately following one of these decisions, because their preceding decision couldn't be confidently categorized as "grant" or "deny".

As in Chen et al. (2016), to restrict the data to instances in which a judge's previous decision is likely to be somewhat salient, I only consider outcomes for which the previous case was decided on the same or previous business day. I consider cases decided on Monday only if the previous case decided by that judge occurred earlier that day, or on the preceding Friday. Given the typically busy case load of an immigration judge, scenarios in which back-to-back cases are separated by a significant amount of time are somewhat rare, and not as relevant to an analysis of the gambler's fallacy.

Critically, even if significant autocorrelation is detected, which indicates that the decision of a previous case affects the decision of the current case, it is not possible to attribute this behavior to a cognitive bias without asserting that there is not corresponding temporal variation in underlying case quality. For example, a decision might have been a grant because the underlying case merit was objectively stronger than the average case seen by the given judge. Looking to the next decision, one would predict a comparatively smaller likelihood of a grant, due simply to a return to average case quality, or regression to the mean. Such judicial behavior would be highly driven by a reaction to change in case quality. It would therefore be a mistake to assign responsibility of corresponding changes in grant behavior to a cognitive bias, rather than to acknowledge it as an appropriate calibration to fluctuating case merits.

Fortunately, cases are randomly assigned to judges within individual immigration courts, using a "first-in-first-out" rule, which means that cases are assigned in the exact order they are received by the court.[81] This allows for the assumption that any temporal variation in case quality originates at the court, rather than the judge level. Episodes such as a surge in refugees from an area that has recently become the site of violent conflict are likely to cause case quality to be positively autocorrelated on the whole, which would bias results against a finding of negative autocorrelation indicative of the gambler's fallacy. Therefore, as in Chen et al. (2016), I control for this time variation in court-level case quality by creating a variable for the number of grants in the last five decisions made by other judges in the same court. In order for this control to accurately reflect recent trends, I restrict the data to observations for which there have been at least five decisions in the same court in the past 30 days. This will control for recent trends in grants, case quality, or judge mood at the court level. To ensure the validity of these court control estimates, I impose Chen et al. (2016)'s restriction of the sample to courts with at least 1,000 associated observations in the data, so that noise in these covariates would be limited.

Research has found that grant rates for particular nationalities vary greatly across different regions and courts.[82] Thus, to control for long term trends by court, I recreated Chen et al. (2016)'s moving variable for the court's average grant rate for the relevant nationality,

[81] Chen, Moskowitz, and Shue, "Decision Making Under the Gambler's Fallacy."

[82] Ramji-Nogales, Schoenholtz, and Schrag, "Refugee Roulette"; U. S. Government Accountability Office, "Asylum"; U. S. Government Accountability Office, "U.S. Asylum System."

excluding the judge at hand. The disparities in grant rates for particular nationalities do not exist only at the court level. Judges within the same court display staggeringly different attitudes toward applicants with the same nationality.[83] Again following Chen et al. (2016), to control for judge-specific tendencies toward specific nationalities, I create a moving variable for the judge's average grant rate for the nationality at hand, which is calculated using all decisions on that nationality previous to, but not including, the current decision.

Finally, I emulate the approach of Chen et al. (2016) and create a series of variables to control for judge-level variation. I first create a variable to control for recent individual judge trends, which simply counts the number of grants in the judge's previous five decisions. This controls for recent changes in a judge's mood or approach to the court. As in the court-level control of last five grants, I restrict the data to observations for which the given judge has made at least five decisions in the past 30 days leading up to the current decision. The use of the moving judge specific control variables (mean nationality-specific grant rates and nationality grants in last five decisions) are advantageous over the use of judge fixed effects because the latter mechanically introduce a slight amount of negative autocorrelation into the model via negative correlation between a judge's previous decision and their current one. To ensure the validity of the judge-specific controls, I impose Chen et al. (2016)'s restriction of the sample to judges who have adjudicated at least 100 cases in the data.

In a divergence from the approach of Chen et al. (2016), I create individual dummy variables for each year of experience up until the 15th year. Cases decided by judges in their 15th or greater year of experience are classified in the same dummy level. To enable a direct comparison of an experience-lagged decision interaction coefficient with that of Chen et al.

[83] Ramji-Nogales, Schoenholtz, and Schrag, "Refugee Roulette."

(2016), I also recreate their “experienced” dummy variable, which indicates that a judge is in their 8th or greater year of experience. For cases with multiple family members, I use only the decision on the lead family member, because almost invariably, all family members receive an identical decision of grant or deny together. I perform this exclusion because the prevalence of unanimity in decisions administered to multiple family members indicates that these should not be considered “separate decisions” in the same sense that other adjacent decisions are considered separate or distinct.

After imposing all restrictions, my data contained 345,083 observations, across 427 judges and 44 immigration courts, which qualified as being fit for use in autocorrelation analysis. A summary of the features of important variables in the data is included in Table 1. It is important to note that while I was successful in replicating most of the controls used in Chen et al. (2016), I was not able to obtain from my data information on whether a given application for asylum was affirmative or defensive. Because evidence suggests that judges treat affirmative and defensive cases differently, with significant implications for grant decisions, my data is missing a very significant covariate.[84] I expect the absence of this variable, which is correlated with the lagged grant decision, as well with the outcome of the current decision, to bias my results against finding negative autocorrelation, because of the nature of how each of these types of applications tend to be processed with respect to time.[85]

[84] Chen, “Explaining Disparities in Asylum Claims.”

[85] Affirmative cases, which are initiated by the refugee, are not expected to be significantly condensed, or bunched in time with respect to their court processing and decision. Defensive cases, however, occur when a non-citizen is apprehended by DHS. Because Immigration Customs and Enforcement (ICE) often conducts “round-ups” in waves, defensive cases are somewhat likely to also be processed by the court in waves, which biases results toward positive autocorrelation.

Summary Statistics: Learning Curve Analysis *(Table 1)*

Statistic	N	Mean	St. Dev.	Median
Decisions	345,083	-	-	-
Judges	427	-	-	-
Courts	44	-	-	-
Lawyer Indicator	-	0.874	-	-
Family Size	-	1.207	0.627	1
Years Since Appointment	-	8.890	6.723	7.567
Morning Indicator	-	0.546	-	-
Lunchtime Indicator	-	0.317	-	-
Afternoon Indicator	-	0.137	-	-
Grant Indicator	-	0.376	-	-
>7 Years Since Appointment Indicator	-	0.475	-	-

Model

I measure evidence of the gambler's fallacy's relationship with judicial experience via a regression using the following linear probability model:

$$Y_{it} = \beta_0 + \beta_1 Y_{i,t-1} + \sum_{j=1}^{14} \beta_{j+1}\, Experience_j + \sum_{j=1}^{14} \beta_{j+15}\, Experience_j * Y_{i,t-1} + Controls + \epsilon_{it} \qquad (1)$$

Y_{it} represents the binary application decision: grant or deny, by judge i ordered by time t.[86] $Experience_j$ represents a set of dummy variables, one of which takes on a value of one when judge i is in their j^{th} year of tenure as an immigration judge. The baseline level of the experience dummy variables is the fifteenth or greater year of experience. The interaction terms can be interpreted as the additional impact of a lagged grant on the probability of current grant for a given experience level, compared to the most experienced judges. A negative coefficient on these terms indicates that less experienced judges are less likely to grant a current decision when

[86] Y_{it} takes on the value of 1 if the current decision by the given judge is a grant. Similarly, $Y_{i,t-1}$takes on a value of 1 if the previous decision by the given judge is a grant.

they have just granted a previous decision than their experienced counterparts. In other words, this would show that inexperienced judges are more severely affected by the gambler's fallacy.

The following variables comprise the controls used in the model: the number of affirmative decisions over the past five decisions (excluding the current decision) of the judge, the number of grant decisions over the past five decisions across other judges (excluding the current judge) in the same court, the judge's mean grant rate for the relevant nationality (excluding the current decision), the court's mean grant rate for the relevant nationality, (excluding the judge associated with the current observation), a presence of lawyer indicator, family size, nationality fixed effects, and time-of-day fixed effects (morning, lunchtime, afternoon).[87] Coefficients can be interpreted as the marginal percentage change in the probability of a grant decision, all else equal. Standard errors are clustered at the judge level to account for the correlation of unobservable characteristics among observations associated with individual judges.[88]

Because case quality for a given judge can be assumed to be independent of time, if the judge's decisions were solely calibrated by this underlying case quality, the coefficient of the β_1 term would be 0, because there would be no reason for certain two-streak decision sequences to more likely than others. Put differently, when controlling for indicators of case quality as well as judge and court trends, a grant in a previous decision should not render the current decision more likely to be either outcome: grant or deny. However, Chen et al. (2016) has already documented the significance of the β_1 term (the effect of a lagged grant rather than deny on the probability of

[87] The inclusion of time of day fixed effects is inspired by research such as that of Danziger et al. 2011, which determined that judicial leniency varied greatly throughout the day, especially in relation to lunch time.

[88] Abadie et al., "When Should You Adjust Standard Errors for Clustering?"; Bertrand, Duflo, and Mullainathan, "How Much Should We Trust Differences-In-Differences Estimates?"

grant), which implied a presence of negative autocorrelation unrelated to case quality in asylum application decisions. Their coefficient on this term ranged between -0.005 to -0.033 across varying specifications and data subsets, indicating that a lagged grant decreased the probability of a current grant by somewhere between 0.5% and -3.3%.[89]

The emphasis of my analysis is on the coefficients of the interaction terms β_{16} through β_{29}, which interact the lagged decision with dummy variables representing the years of tenure accrued by the given judge at the time of a decision. Again, these interaction coefficients represent the marginal effect that a given experience level has on the impact of a lagged grant on the probability of granting asylum in the current decision. As theory indicates that committal of cognitive errors becomes less frequent as experience is accrued, I not only expect these interaction terms to be significant, but also to demonstrate a reduction in impact of the lagged decision on the current decision as judicial experience increases.

Results

As discussed in the data section, I anticipated that the inability to acquire a variable representative of the defensive status of asylum applicants would prove problematic regarding the estimation of the main effect lagged decision term. I determined that omitted variable bias was likely to exert an upward bias on this estimate, which would, at best, cause my main effect coefficient to be slightly less negative than those estimated by Chen et al. (2016), and at worst, qualitatively change the coefficient to indicate that a previous grant *increases* the likelihood of a grant, which contradicts Chen et al. (2016), and is inconsistent with evidence of the gambler's fallacy. It is critical to note that the omitted variable bias which alters my main effect does not

[89] Chen, Moskowitz, and Shue, "Decision Making Under the Gambler's Fallacy."

compromise the validity of the interaction effects in the same regression model. Several econometric papers have documented the consistency of interaction effect estimation despite the presence of omitted variable bias.[90] In cases where treatment (the lagged grant indicator) is endogenous due to omitted variables, and the source of heterogeneity (judicial experience) is exogenous, the ordinary least squares estimator of the coefficient of the interaction term remains consistent, despite a potentially biased main effect. Indeed, my lagged grant indicator is endogenous, while my judicial experience indicators are exogenous.

Ultimately, though my main effect estimate is unreliable, the coefficients of interest and focus of this research can still be estimated consistently with confidence.[91] Furthermore, the interpretation of these coefficients with respect to the theory of a learning curve does not change so long as the true main effect of a lagged grant is negative (as was found to be true across multiple model variations by Chen et al. (2016). Therefore, instead of using a biased main effect to ground an analysis of the interaction terms, I proceed by accepting Chen et al. (2016)'s findings as ground truth: that there exists a small, but significant and robust negative effect of a previous grant on the probability of a current grant. The exact value of this estimate is consequential only for estimating the precise magnitude of autocorrelation interacted with experience, which is not the concern of my inquiry. Instead, as emphasized previously, the focus of my research is on the relative degree of autocorrelation in decisions exhibited by judges at different experience levels.

In Table 2, I present results for interaction terms between experience and the lagged decision from the regression model outlined by Equation 1 (Individual Year Model), as well as

[90] Bun and Harrison, "OLS and IV Estimation of Regression Models Including Endogenous Interaction Terms.", Nizalova and Murtazashvili, "Exogenous Treatment and Endogenous Factors."

[91] The main effect of a lagged grant estimated in my model was 0.07, which is inconsistent with the results of Chen et al. 2016. I attribute this result to omitted variable bias, and consider my interaction terms independent of this result.

from a model in which the set of interactions involving dummies for each year of experience are replaced by a single interaction term between a dummy indicating whether a judge has over seven years of experience, and the lagged decision (Binary Experience Model):[92]

Autocorrelation Results *(Table 2)*

	Grant	
	Individual Year Model	Binary Experience Model
Lag Grant:1st Year	-0.073*** (0.010)	-
Lag Grant:2nd Year	-0.054*** (0.009)	-
Lag Grant:3rd Year	-0.047*** (0.010)	-
Lag Grant:4th Year	-0.035*** (0.009)	-
Lag Grant:5th Year	-0.041*** (0.010)	-
Lag Grant:6th Year	-0.036*** (0.010)	-
Lag Grant:7th Year	-0.032*** (0.011)	-
Lag Grant:8th Year	-0.040*** (0.011)	-
Lag Grant:9th Year	-0.020*** (0.010)	-
Lag Grant:10th Year	-0.022*** (0.011)	-
Lag Grant:11th Year	-0.001 (0.011)	-
Lag Grant:12th Year	-0.0002 (0.011)	-
Lag Grant:13th Year	-0.016 (0.013)	-
Lag Grant:14th Year	0.015 (0.012)	-
Lag Grant:15th Year +	-	-
Lag Grant: Experienced Dummy (>7 Years)	-	0.039*** (0.003)
N	345,083	345,083
R^2	0.374	0.373
Adjusted R^2	0.374	0.373
Residual Std. Error	0.383 (df = 344870)	0.383 (df = 344894)
F Statistic	973.054*** (df = 212; 344870)	1,107.570*** (df = 188; 344894)

Notes: "Nth" year refers to a judge having between n-1 and n years of experience adjudicating asylum cases.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

I find, predicated on the assumption that the true main effect is small and negative, that a lagged grant decreases the probability of a grant much more significantly for judges in their earlier years of tenure than for those in their later years. The coefficients indicate that in

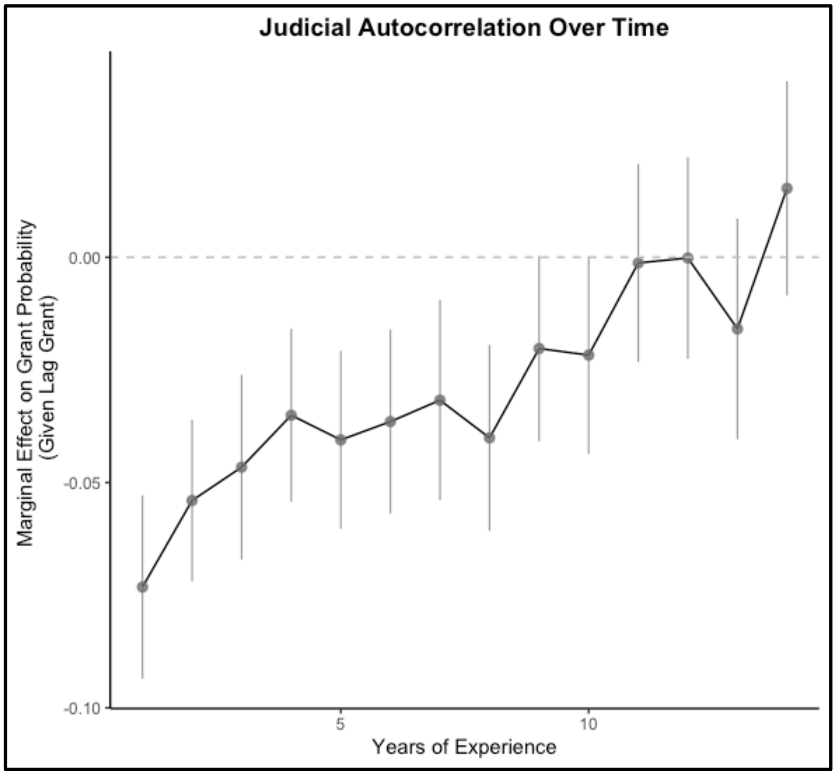
[92] When the individual year experience dummies in the interaction were condensed to a single binary dummy indicating whether a judge had more than 7 years of experience, the experience control variable was similarly condensed into a dummy with only two levels.

decisions following a grant, a judge in her first year of experience is 7.3% less likely to grant the asylum application in their current case than a judge in their fifteenth year. In other words, the gambler's fallacy effect appears to be *much* stronger when judges lack experience. Each interaction term up to and including the tenth year of experience is not only negative, which indicates a greater exhibition of gambler's fallacy-type behavior than that of the most experienced group, but also statistically significant with a p-value of less than 0.01. The effect sizes generally trend to zero as experience increases, and indicate that after ten years of tenure, the average judge accrues no additional benefit (in terms of resilience to the cognitive biases responsible for the gambler's fallacy) from an extra year on the bench. This is reflected by the coefficients Lag Grant:11th year through Lag Grant:14th year not differing significantly from the baseline interaction effect of zero.

I included the Binary Experience Model, which was identical to my first model other than the substitution of the yearly experience levels for that of a binary experience term. The coefficient of interest was therefore the interaction between this experience dummy and the lagged decision term. I confirmed that coefficient of this term was indeed consistent with the findings of Chen et al. (2016). Indeed, despite the disparity between my paper and Chen et al. (2016) in main effect estimates due to omitted variable bias, my reported estimate for the interaction is 0.039. This is squarely in-between Chen et al. (2016)'s reported coefficients of 0.033 and 0.046 (produced using slightly different sub-samples of data). This alignment further confirms the stand-alone validity of my interaction term coefficients.

Figure 1, which displays the interaction coefficients plotted against their corresponding year of experience, is effective in clarifying the existence and shape of the "learning curve", at least with regard to the cognitive biases that give rise to the gambler's fallacy.

Figure 1: Gambler's Fallacy Learning Curve



Error bars above and below each point represent a 95% confidence interval around the estimate. The curve rises relatively sharply and linearly in the first four years of experience, indicating a quick initial improvement. As discussed, after year ten, the interaction effect does not differ significantly from zero (evidenced by the intersection of the error bars with the x-axis), meaning that it no longer is different from that exhibited by the oldest experience cohort, who are in their fifteenth or greater year of experience. It is important to note that just because these interaction terms are indistinguishable from zero does not mean that judges with more than ten years exhibit no cognitive bias at all – it simply means that these judges exhibit as little cognitive bias as they ever will. Thus, the learning curve exhibits significant diminishing marginal returns with respect to gains in resilience to the cognitive bias of the gambler's fallacy.

To evaluate the robustness of the tenth year's significance, I also run logit and probit regressions, the results of which are included in the appendix (Tables 8 and 9 respectively). In both of these models, the tenth year is the last to exhibit an interaction effect significantly different from zero. To assess the stability of the estimates across different experience groupings within the baseline dummy level, I vary this baseline level so that it takes on the following categorizations: >10 years, >11 years, >12 years, >13 years, >15 years, >16 years. For each categorization, I run the same linear probability model outlined in Equation 2. The results of these models are also included in the appendix (Table 10), and all corroborate the finding that the tenth year is the last in which a judge can expect to improve in their susceptibility to the gambler's fallacy.

Due to the apparent relevance of the tenth year of experience as the final year in which a judge improves in their decision maturity, or resilience to cognitive bias, I use the ten-year mark to divide "experienced" and "inexperienced" judges in my discrimination analysis to follow. I categorize judges with greater than ten years of tenure as experienced, and judges with ten or less years as inexperienced. This produces a meaningful dichotomy, in which the experienced judges group are confirmed to have reached a type of cognitive maturity, while the inexperienced judges are still demonstrating significantly more susceptibility to cognitive bias than their experienced counterparts.

Post-9/11 Discrimination

In this section, I detail the data and methods used in, as well as the results from, my analysis of the relationship between judicial experience and anti-Muslim discrimination following 9/11. I perform this analysis by making use of the natural dichotomy that appears to exist in cognitive maturity suggested by the learning curve constructed in the previous section.

By deliberately creating experience with respect to each group's susceptibility to cognitive bias, I am able to evaluate whether this cognitive immaturity corresponds to discriminatory behavior.

Data

In analyzing post-9/11 anti-Muslim discrimination among judges of different experience levels, I use much of the same asylum data detailed in the previous section. Unlike the learning curve exercise, observations used in this phase of analysis are not dependent on the validity of the observations immediately prior. Therefore, while I maintained the exclusion of application decisions labeled "other", which could not confidently be categorized, I was able to retain observations which themselves were clearly grant or deny, but were preceded by an "other".

Also in contrast to the learning curve analysis, I was also able to retain observations which previously couldn't be placed on a known timeline because a date was recorded, but a specific time was not. I restrict the data to decisions that occurred between December of 1996 and December of 2002, as I am interested primarily in a shorter-term behavioral response to 9/11, rather than one that persisted for multiple years. The value of analyzing data beyond December 2002 is further diminished by the start of the Iraq War in early 2003, which poses difficulties for isolating the effects of the 9/11 attacks on discrimination against Middle Easterners.

Ideally, the asylum data would contain a variable indicating the religion of each applicant. Information this granular would allow me to very clearly identify a decline in grant rates among those who identify as Muslim following 9/11. Since I do not have access to such a variable, I am forced to rely on a blunter instrument, using the applicant's country of origin as a sort of proxy for negative effects befalling Muslim asylum applicants after 9/11. Assuming that the composition of asylum applicants hailing from Muslim-majority countries is decently reflective of these countries' religious populations, this strategy should not prove significantly detrimental to my ability to detect the effects of anti-Muslim discrimination. If anything, I expect the use of the proxy Muslim variable to bias my results against finding an effect, because the treatment group will be diluted with non-Muslim applicants, who should be less likely to receive different treatment following 9/11. Such bias could be partially counteracted by the human tendency to stereotype, in which judges group applicants based on their country of origin, and therefore project anti-Muslim bias (subconscious or not) on all applicants from Muslim-majority countries, rather than at a more individual level.

Because I am focused on macro-level phenomena, e.g. how certain nationalities of asylum seekers were affected by 9/11, rather than individual case-level decisions, the outcome of interest in this phase of analysis is the average grant rate for asylum applications filed by members of a treatment group, compared to the average grant rate for applications filed by members of other "control" nationalities. Given the wave of anti-Muslim sentiment perpetuated by 9/11, my focus for the treatment group is primarily on the nationalities that were most likely to become the targets of such discrimination.[93] This particular strategy of identifying treatment groups has been previously leveraged in studies on the effects of terrorist attacks on Muslims in immigration court.[94] Following Rabby and Rodgers (2011), I restrict my treatment group to Middle Eastern Arab countries, whom research shows had the greatest likelihood of being associated with terrorist stereotypes in the American consciousness.[95] The final treatment group includes applicants whose stated country of origin is one of the following: Bahrain, Iran, Iraq,

[93] Rabby and Rodgers, "Post 9-11 U.S. Muslim Labor Market Outcomes."; Though many countries such as Indonesia, Pakistan, or Algeria are also Muslim-majority, these groups were not as closely associated by the American public with terrorist stereotypes following 9/11. This is potentially due somewhat to their locations outside of the Middle-East.

[94] Mensah and Opoku-Agyemang, "Innocent Until Stereotyped Guilty? Terrorism and US Immigration Court Decisions."

[95] Rabby and Rodgers, "Post 9-11 U.S. Muslim Labor Market Outcomes"; Mensah and Opoku-Agyemang, "Innocent Until Stereotyped Guilty? Terrorism and US Immigration Court Decisions."

Jordan, Kuwait, Lebanon, Oman, Palestine, Qatar, Saudi Arabia, Syria, Yemen, United Arab Emirates, and Afghanistan.[96] All other countries are eligible to be used as controls.

The data is split into two segments: application decisions made by "experienced" judges, and those made by "inexperienced" judges. The division between the two groups is determined by the learning curve cut-off of ten years of tenure. The data is aggregated in both segments so that a grant rate is calculated for each nationality per fiscal quarter.[97] To balance the goal of reducing noisiness and volatility in the data with that of maintaining a robust collection of control units, I discarded from the sample nationalities that did not have at least five asylum applications adjudicated per quarter. A requirement of any more than five asylum applications per quarter would have significantly decreased the countries eligible for analysis.

Because the median judge experience over the time period of interest is only 5.2 years, there are far fewer decisions adjudicated by "experienced" judges (those with more than ten years of tenure) in the data. There are only 65 judges in the experienced group, as compared to 198 in the inexperienced group. The limitation of nationalities in each experience group to those who had five observations per quarter largely explains the relatively smaller number of control countries in the experienced cohort. It also explains the fact that each data segment has different control countries– a country that is listed as a control in the inexperienced group, but not in the experienced group can be presumed not to have fulfilled the quarterly data requirement in the latter. Covariates were not used in this phase of analysis; therefore the data can be characterized

[96] Cooley, "The Contagion Spreads."; While Afghanistan is not technically part of the middle-east, it is the only geographical exception on in the treatment group. It warranted treatment status because of its history of conflict with the United States preceding 9/11, and likelihood to be associated with jihadist terrorism, largely due to the presence and influence of the Taliban. Additionally, one of the terrorists in the 9/11 attacks was Afghan.

[97] While observations are grouped in intervals that reflect the length of fiscal quarters, they are adjusted to avoid contamination of the data or treatment effect. Accordingly, the quarter before September 11th, 2001 ends on September 10th 2001, and the following quarter begins on September 11th.

as panel data consisting simply of quarterly average grant rates for each control nationality and the treatment group.

To meaningfully compare grant rates from before and after 9/11 for applicants from Middle Eastern Arab countries, I must verify that the distribution of case quality for this treatment group did not change significantly between the two time periods. If the cases filed after 9/11 had, for whatever reason, inferior claims to asylum than those filed before, a drop in grant rates would be expected in the absence of anti-Muslim discrimination. I determine that a change in the composition of treatment group applicants is not likely to have occurred for two reasons.

First, the backlog of the court ensures that wait times for asylum applicants (time between the scheduling of their Merits Hearing and the actual Merits Hearing) are frequently between two and four *years*.[98] Much of my data post 9/11 therefore is composed of applicants who filed well before the attacks, and would not reflect any compositional changes that may have occurred after the fact. I also sought to verify this claim by using the distribution of countries of origin as a proxy for case quality, as I do not have access to better underlying measures of case quality. I conducted a comparison of the proportion of the treatment group represented by each country before and after the attacks, which can be found in the appendix (Table 11, Figure 13).[99] I find little difference in the makeup of treatment group applicants.[100]

The largest change in the comparative distributions is driven by an increase in Iraqi applicants, who made up 17.4% of the treatment group before the attacks, and 26.2% after. Importantly, this influx did not disproportionately affect one experience group – the groups each

[98] "Despite Hiring, Immigration Court Backlog and Wait Times Climb."

[99] In Table 11, the cells without values signal the data contains no applications for the given nationality in the indicated time period. In Figure 13, nationalities are omitted from both pie charts if they do not account for at least 1% of the treatment group in both periods.

[100] I restrict the time periods analyzed before the attacks to the five quarters preceding the attacks, rather than use all nineteen quarters of data leading up to the attacks. This is to match the five periods of data I use after the attacks so as to reflect similar amounts of noise and volatility, as well as more recent trends.

experience a similar change in the proportion of the treatment group made up by Iraqi applicants.[101] I calculate the grant rate for Iraqi applicants among each experience group before and after the attacks, and find that while the grant rate for inexperienced judges does not change by more than 1%, the grant rate for experienced judges drops by roughly 7%. This drop could easily be due to the small sample size of experienced group, but to be conservative, I consider the increased proportion of Iraqi cases (which may be of worse average quality) as a possible bias which could misleadingly give the appearance evidence of discrimination in the experienced group.[102] Otherwise, I find scant evidence that a significant shift in the distribution of case quality threatens my ability to identify anti-Muslim discrimination.

Ramji-Nogales et al. (2007) find that the political party of the president whose administration appointed an immigration judge is not a significant predictor of judge decision patterns.[103] However, because political affiliation played a significant role in the public reaction to 9/11, particularly with respect to increased prioritization of national security concerns, I investigated the distribution of judges across experience groups with respect to whether they were appointed by Democrat or Republican administrations.[104] I conduct a chi-squared test, the results of which are in the appendix (Table 12), and do find a significant difference in the (somewhat presumed) political makeup of the experience groups.[105] Roughly 72% of inexperienced judges were appointed by Democratic administrations, as compared to only about 15% of experienced judges.

[101] More specifically, before 9/11, Iraqi applicants made up 17.2% of all treatment group cases for inexperienced judges, and 17.8% for experienced. After 9/11, Iraqi applicants made up 26.6% of all treatment group cases for inexperienced judges, and 25.4% for experienced.

[102] The experienced judges adjudicated only 68 cases before the attacks, and 102 afterwards. It's very possible that the 7% drop was not a function of material changes in attitude toward the average Iraqi applicant, but rather a function of a volatile grant rate due to small sample size.

[103] Ramji-Nogales, Schoenholtz, and Schrag, "Refugee Roulette."

[104] Skitka et al., "Confrontational and Preventative Policy Responses to Terrorism: Anger Wants a Fight and Fear Wants 'Them' to Go Away."

[105] I say "somewhat presumed" because although more restrictive grant patterns tend to be correlated with judges appointed by Republican administrations, not every judge is politically aligned with the administrated which appointed them, and in many cases the political leanings of immigration judges aren't officially known by the government.

Based on the findings of Ramji-Nogales, this lopsided distribution should not impact my results. If, however, Republican-appointed judges do exhibit more stringent grant behavior reflective of elevated national security concerns following 9/11, I would expect the experienced judge estimate to be biased downward, exhibiting a sharper drop in grant rates than the true estimate. Conversely, I would expect the inexperienced estimate to be biased upward, exhibiting a less severe drop in grant rates than the true estimate. In short, because of an imbalance between experience groups, politicized judge behavior would cause results to be biased against my hypothesis that inexperienced judges should display a sharper drop in grant rates than experienced judges. Table 3 contains summary information for the final versions of the experienced and inexperienced data sets respectively. The similarity in mean grant rates between the two groups is notable. That experienced and inexperienced judges approve effectively (a less than 1% difference) the same portion of all cases supports the notion that their decisions are appropriate for comparison in this analysis. This implies that differences in behavior following 9/11 are more easily attributed to the effect of cognitive biases rather than systematic differences between the experience groups.

Summary Statistics: Post-9/11 Discrimination Analysis *(Table 3)*

Group	Statistic	N	Mean	St. Dev.	Median
Experienced	Countries*	32	-	-	-
	Quarterly Grant Rate	792	0.344	0.192	0.334
Inexperienced	Countries*	56	-	-	-
	Quarterly Grant Rate	1,368	0.353	0.191	0.339

**Note: Countries counted excluding those part of treatment group*

Model

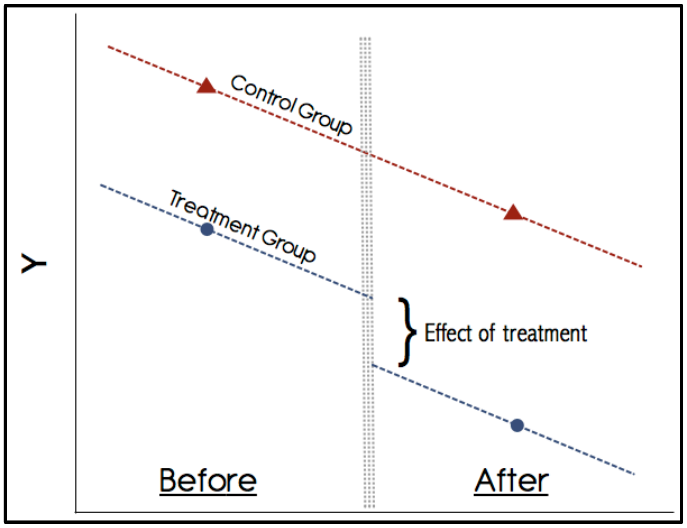
The effect of 9/11 on the decisions rendered by experienced and inexperienced judges on applicants from Middle Eastern Arab countries might initially seem a trivial matter of comparing grant rates before and after the attacks for each experience group. In order to actually estimate the size of 9/11's effect on anti-Muslim discrimination, however, I must distinguish between the treatment group's "untreated" counterfactual, and the treatment group's observations before the attack, in the "pre-period".

To calculate the true effect, I must compare the observed judicial behavior, given the occurrence of the attack, with the hypothetical judicial behavior that would have occurred in the absence of such an event. Because I expect judges to exhibit some degree of natural variation in their decision patterns, it is important to ensure that behavioral changes observed are indeed due to the effects of 9/11, and would not simply have happened in absence of the attack. A naïve comparison of "pre" and "post" 9/11 judicial behavior toward the treatment group of Middle Eastern Arab countries invokes the strong assumption that the behavior in the pre-period accurately, or at least satisfactorily approximates, the hypothetical behavior in the post-period that one would have expected to observe if the attacks did not actually happen. Obviously, this counterfactual behavior for the treated group is impossible to empirically observe, because there is simply no world in which 9/11 did not happen, and anti-Muslim discrimination did not ensue.

The next best option, then, involves leveraging observations of asylum seekers from a "control" country, or group of countries. These control observations can be presumed not to have been affected by discrimination following 9/11 in the same manner as the treatment group. If the trends in the control units are similar to that of the treatment unit prior to 9/11, the control units, which were not affected by the impact of treatment, can be used to approximate the trajectory of the treatment into the post-period, thereby allowing for the construction of a counterfactual. Importantly, for this approach to yield reliable treatment effect estimates, the control countries' data need not reflect similar grant rates to those of the treatment group.

Rather, so long as the control and treatment pre-period data exhibits "parallel trends", a simple difference between the two groups can be taken in the pre-period, and compared to the difference between the groups in the post-period. As illustrated in Figure 2, the size of the difference between control and treatment changed after the treatment was effected. This change in differences is typically interpreted as evidence of a treatment effect. Such is the logic behind the popular causal identification technique of "Difference in Differences".

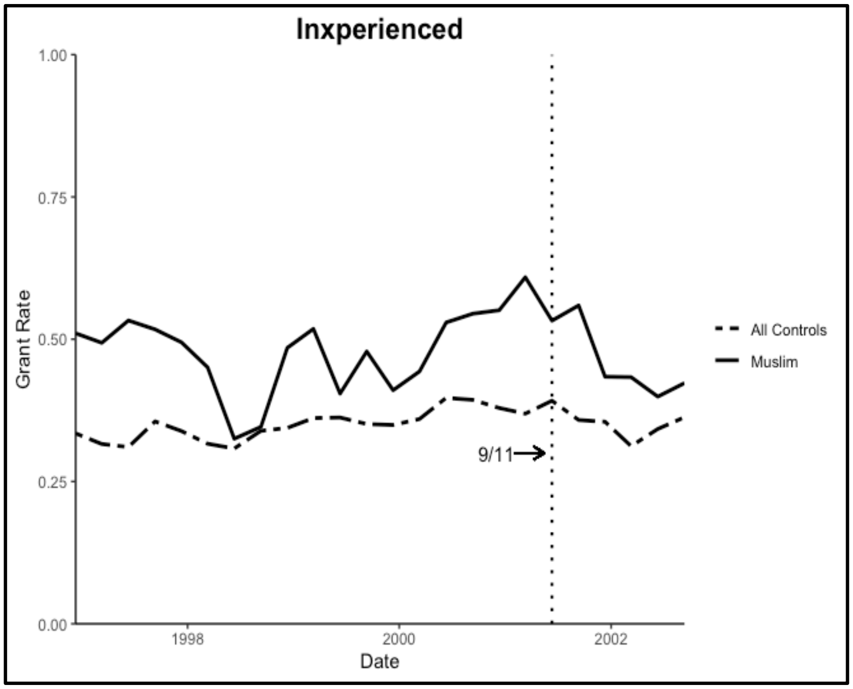
Figure 2: Difference in Difference Example



It is not clear however, that a Difference in Differences model is appropriate for application in my context. First comes the question of which country or group of countries should be used as control units for comparison with the treatment group of Middle Eastern Arab countries. Second, the asylum grant data exhibits significant volatility, making the identifying assumption of parallel trends extremely difficult to satisfy, or even verify. Further complicating the satisfaction of this assumption is the focus of my analysis on heterogeneity across experienced and inexperienced judges. Comparing Difference in Difference estimates across models raises questions about the need to satisfy parallel trends not only within each experience cohort, but across cohorts as well, meaning that the pre-trend control group for experienced judges must be reasonably parallel to the pre-trend treatment group for inexperienced judges, etc.[106]

If one considers observations from all countries not designated as "treatment" in the data to act as controls, it is clear from Figures 3 and 4 that the average grant rate trends for the treatment and all control units struggle to satisfy the parallel trends assumption for either experience group.[107] As a logistical note, because the formation of the treatment group was based on an anticipation of anti-Muslim discrimination, the treatment group is labelled "Muslim" rather than "Middle Eastern Arab" for brevity.

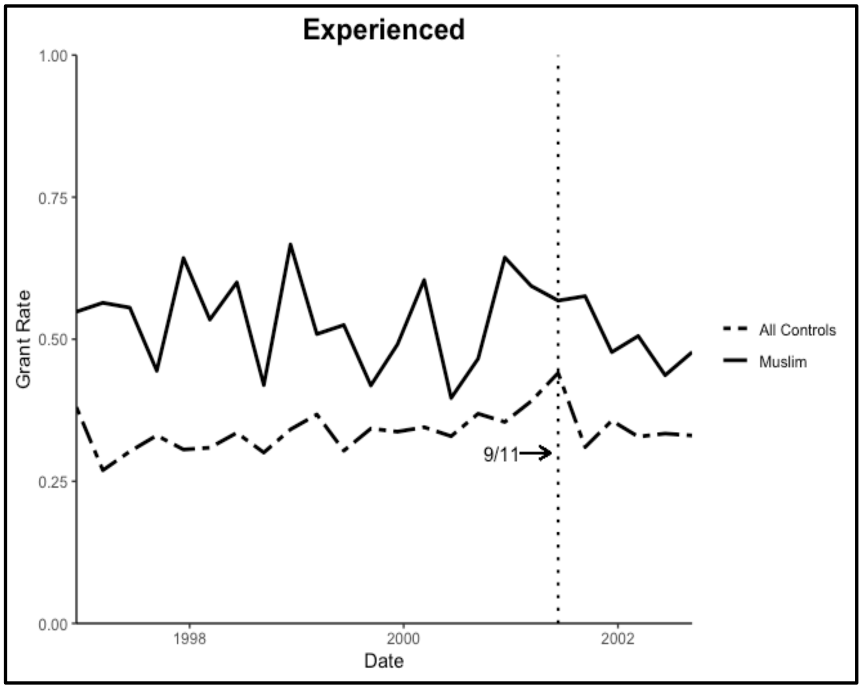
Figure 3: Treatment vs All Control Grant Rate Trends (Inexperienced)



[106] Basant and Sen, "Impact of Affirmative Action in Higher Education for the Other Backward Classes in India."

[107] The parallel trends assumption cannot be empirically verified due to the absence of the treatment unit's counterfactual in the post-period. Thus, parallel trends in the period prior to the treatment are typically accepted as a de facto satisfaction of the parallel trends assumption.

Figure 4: Treatment Vs All Control Grant Rate Trends (Experienced)



The shortcomings of a Difference in Difference approach in this context motivate the use of an alternative, non-parametric method: synthetic control, which has gained significant popularity in social science literature in recent years. It is particularly well-suited to research contexts such as mine, in which there is effectively a single treatment unit, and many possible control units. Synthetic control is especially advantageous in that it solves the problem of which units to leverage as "controls" against a treatment, by constructing its own novel control unit from linear combinations of existing possible controls. This allows a treatment unit (such as applicants from Middle Eastern Arab countries), which may seem incomparable on one or several dimensions to the available control units, to receive a closer "matched" control, via a synthetic unit than would otherwise be possible.

The model assigns various weights to each control unit as it attempts to minimize the difference between its synthetic control and the treatment unit throughout the pre-period. In other words, countries in the control group are each designated a certain degree of influence on the final synthetic control, so that the combination of weights chosen produces a control unit which approximates the pre-period treatment unit as best as possible.

More formally, the total number of time periods (quarters) in the sample can be represented by $T = T_B + T_A$, where there are T_B periods *before* 9/11, and T_A periods *after* 9/11. The number of control countries can be denoted by I. Let $Y_1 = \begin{bmatrix} Y_{1B} \\ Y_{1A} \end{bmatrix}$ denote the $Tx1$vector of average grant rates for the treatment group of asylum applicants from Middle Eastern Arab countries, where Y_{1B} is the T_Bx1 vector of pre-9/11 average grant rates, and Y_{1A} is the T_Ax1 vector of post-9/11 average grant rates. Similarly, let $Y_0 = \begin{bmatrix} Y_{0B} \\ Y_{0A} \end{bmatrix}$ denote the TxI matrix of average grant rates for the control countries, where Y_{0B} is the T_BxI matrix of pre-9/11 average grant rates, and Y_{0A} is the T_AxI matrix of post-9/11 average grant rates. Given that W is an $Ix1$ vector of weights, the synthetic control country is chosen by selecting weights W to solve

$$argmin(Y_{1B} - Y_{0B}W)'V(Y_{1B} - Y_{0B}W) \qquad (2)$$

subject to the constraints that all weights be non-negative, and add up to a total of exactly one. V represents a diagonal, non-negative matrix, and is able to place more importance on some observations in time than others. The estimated treatment effect of 9/11 on average grant rates for applicants from Middle Eastern Arab countries, is given by:

$$\hat{\Delta} = \frac{1}{T_A} \sum_{t=T_{B+1}}^{T} (Y_{1At} - Y_{1At}^{*}) \qquad (3)$$

where Y_{1At} (the observed mean grant rate for the treatment group in post period t) is a given element of Y_{1A} (a matrix of all post-period mean grant rates for the treatment group), where Y_{1At}^{*} (the observed mean grant rate for the synthetic control in post period t) is a given element of $Y_{1A}^{*} = Y_{0A}W^{*}$ (a matrix of all post-period mean grant rates for the synthetic control), and where W^* is a vector of the weights acquired by solving Equation 2. This process can ultimately be thought of as the creation of a weighted average of control units in which weights are assigned to control units so as to best imitate the pre-treatment behavior of the treatment unit, and some countries can be designated a weight of zero if they are not useful enough in this approximation to be included in the weighting scheme.

Notably, it is possible to alter the matching process so that more importance is placed on minimizing the difference between the synthetic control and the treatment unit in later stages of the pre-period, closer to the intervention. In cases such as asylum, where grant rate data is significantly volatile, this strategy of incurring more losses in the earlier pre-periods in return for greater accuracy in the later pre-periods may be preferred so that the synthetic control tracks more closely directly prior to 9/11, and resulting gaps in the post-period aren't as likely to stem from a consistently poor fit rather than an actual treatment effect. In my synthetic control models for both the experienced and inexperienced group, (which contain 19 pre-period quarters) therefore, I choose to optimize the matching process on quarters 6 though 19, which comprises slightly more than two-thirds of the pre-9/11 data.

As discussed earlier, to reflect a focus on a short-term behavioral response, as well as avoid contamination of treatment effects via the Iraq War, I evaluate post-9/11 behavior for only five quarters, through December of 2002.[108] Conveniently, time period-specific treatment effects can be obtained by taking the difference between the synthetic control and the treated unit in a given quarter following 9/11. Otherwise, as shown in Equation 3, all post-9/11 treatment effects can be averaged to obtain an average treatment effect for the entire post-period.

[108] Post-period quarters are represented by the following five date intervals: 09/11/01 – 12/10/01, 12/11/01 – 03/10/02, 03/11/02 – 06/10/02, 06/11/02 – 09/10/02, 09/11/02 – 12/10/02.

In some synthetic control applications, the synthetic comparison unit is constructed by matching not only on the outcome variable, but also on other covariates predictive of the outcome.[109] This can be helpful in certain settings to ensure that the synthetic control resembles the treatment units on several dimensions. The volatility of the asylum data, however, which is exacerbated in my sample by a smaller number of observations in the experienced cohort, means that the use of additional covariates in the construction of the synthetic comparison unit is unlikely to add value to the matching process, because of the inconsistency of covariate values within individual units over time. Therefore, as in Grogger (2017), I allow the synthetic control to match only on quarterly observations of the outcome variable (average grant rate), and do not use covariates.[110]

To assess the significance of treatment effect estimates produced by the synthetic control, I follow Abadie et al. (2010), borrowing from the logic of permutation inference to create a distribution of treatment effects under the null hypothesis of no treatment effect.[111] Each control country takes a turn acting as the "placebo" treatment group, and a corresponding synthetic comparison unit is created from the other control countries via the matching and optimization process detailed above. I then calculated placebo treatment effects for each of the control units.

Because presumably the control units were not subject to the same anti-Muslim discrimination as the treatment group, I do not expect 9/11 to trigger a change in the immigration judges' treatment of the control applicants. The post-9/11 behavior of the control units should not, in theory, deviate much from their corresponding synthetic comparison units, which were trained on the pre-9/11 data. The distribution of these placebo treatment effects should be

[109] Abadie, Diamond, and Hainmueller, "Synthetic Control Methods for Comparative Case Studies."

[110] Grogger, "Soda Taxes and the Prices of Sodas and Other Drinks."

[111] Abadie, Diamond, and Hainmueller, "Synthetic Control Methods for Comparative Case Studies."

centered more or less around zero. Creating this distribution against which to compare the true estimated treatment effect enables me to assess whether this treatment effect is large relative to the effect estimated for a control country chosen at random. Under a null hypothesis of no treatment effect, the true estimated treatment effect obtained by the synthetic control would not be abnormal relative to the distribution of the placebo treatment effects.

The relation of the observed treatment effect compared to the placebo effects can be quantified in the synthetic control version of a p-value by dividing the number of placebo units that have treatment effects at least as extreme than the observed treatment effect by the total number of placebo units.[112] Especially for a study with many control units, this calculation is meant to roughly represent the probability of obtaining test results at least as extreme as the actual results, assuming that the null hypothesis is correct.

Nowhere in the synthetic control methodology is the synthetic comparison unit's fit to the treatment unit in the pre-treatment period guaranteed to be perfect, or even good. A synthetic comparison unit that very poorly tracks the treatment unit may produce unreliable treatment effects due to its inability to model the trajectory of the treatment unit into the post-treatment period. A statistic called MSPE (pre-treatment mean squared prediction error) can be helpful in assessing the quality of a synthetic comparison unit. In my context, the MSPE represents the average of the squared differences between true average grant rates in the treatment group and in its synthetic counterpart up until 9/11. Lower MSPE values correspond to a better synthetic control fit. It is common to restrict analysis of the placebo distribution to placebo controls that

[112] The term p-value is used loosely here. In reality, the method described can fail to precisely mirror the properties of the p-value in large sample parametric models due to a small control pool, or a choice of restriction by pre-treatment MSPE regarding which units are considered viable members of the placebo distribution.

meet a specific threshold of MSPE, so that the distribution is comprised only of treatment effects produced with a certain degree of confidence and fit.

Another helpful statistic for interpreting the obtained treatment effect in relation to the placebo effects is the post/pre-treatment MSPE ratio. This value is obtained by dividing the post-treatment mean squared prediction error by the pre-treatment mean squared prediction error.[113] The MSPE ratio is informative because in the event of a significant treatment effect, the post-treatment MSPE will be high, and in the event of a high-quality synthetic control (one that closely approximates the treatment unit's pre-treatment behavior), the pre-treatment MSPE will be low. A higher MSPE ratio therefore reflects a large deviation of the treatment unit from the synthetic unit in the post-period, as compared to the deviation that existed prior to treatment.

Rank-ordering the MSPE ratio for the treatment among the MSPE ratios observed for placebo units provides additional information regarding how the treatment's post-period deviation compares to the deviations produced under the hypothesis of no treatment effect. The estimated treatment effect ranking very high on this list supports the rejection of this null hypothesis. Ultimately, the significance of a synthetic control treatment effect is best evaluated not by a single statistic, but by thoughtfully synthesizing information from the placebo distribution (possibly under pre-MSPE restrictions), the associated "p-value", and the post/pre-treatment MSPE ratio. This synthesis should be accompanied by a consideration of the influence of sample size on these measures.

Because I am primarily focused on the *difference* in discriminatory reactions to 9/11 between experienced and inexperienced judges, I perform the synthetic control process separately on each experience group's data. While there currently exists no formal method of

[113] Abadie, Diamond, and Hainmueller, "Synthetic Control Methods for Comparative Case Studies."

statistical inference for comparing results from separate synthetic controls, useful comparisons can still be made between the two models to facilitate findings regarding heterogeneity. Comparisons of treatment effects obtained from separate synthetic control models have been conducted in several economic research papers, including Jones and Marinescu (2018).[114]

Results

In this section I present results from synthetic control models regarding changes in asylum grant rate behavior following 9/11 for inexperienced and experienced immigration judges. As discussed, a synthetic control, in my research context, is essentially a weighted average of select control countries' grant rates. The synthetic control forms a novel comparison unit that aims to mirror the grant rates of the Muslim treatment group (Middle Eastern Arab applicants) prior to 9/11. In the period following 9/11, the difference between the synthetic control grant rate and the grant rate for the Muslim treatment group represents the effect of the attacks on grant rates for the treatment group. I apply the synthetic control method to data associated with inexperienced and experienced judges separately.

Results for the experienced and inexperienced group are displayed in Table 4 and 5 respectively. These tables compare the pre-9/11 average grant rates of the treatment group with those of the synthetic treatment, as well as with the average of all countries in the control pool. It is apparent that, for both inexperienced and experienced groups, the values in the synthetic column are much closer to those of the treated group than the values in the average of all controls column. This is further evidence that the average of all controls does not provide a satisfactory comparison group, and confirms the advantages of the synthetic control method.

[114] Jones and Marinescu, "The Labor Market Impacts of Universal and Permanent Cash Transfers."; Jones and Marinescu create and compare separate synthetic controls to evaluate heterogeneity across marriage status and gender in labor force participation following the receipt of cash transfers.

Grant Rate Means: Inexperienced *(Table 4)*			
Quarter*	Treated	Synthetic	Average of 56 Control Countries
1996-12-11	0.510	0.383	0.335
1997-03-11	0.494	0.526	0.316
1997-06-11	0.533	0.413	0.311
1997-09-11	0.517	0.473	0.356
1997-12-11	0.495	0.577	0.339
1998-03-11	0.451	0.441	0.316
1998-06-11	0.325	0.333	0.308
1998-09-11	0.346	0.367	0.339
1998-12-11	0.485	0.476	0.344
1999-03-11	0.518	0.499	0.361
1999-06-11	0.404	0.422	0.362
1999-09-11	0.478	0.480	0.351
1999-12-11	0.410	0.423	0.349
2000-03-11	0.443	0.435	0.360
2000-06-11	0.529	0.547	0.397
2000-09-11	0.545	0.541	0.393
2000-12-11	0.551	0.549	0.379
2001-03-11	0.609	0.583	0.369
2001-06-11	0.533	0.536	0.392

**Note: Quarters reported using a "floor date", e.g. any date including or after 1996-12-11 & prior to 1997-03-11 is classified as belonging to quarter "1996-12-11"*

Grant Rate Means: Experienced *(Table 5)*			
Quarter*	Treated	Synthetic	Average of 32 Control Countries
1996-12-11	0.548	0.373	0.380
1997-03-11	0.564	0.304	0.269
1997-06-11	0.556	0.392	0.303
1997-09-11	0.444	0.380	0.331
1997-12-11	0.643	0.417	0.306
1998-03-11	0.534	0.487	0.309
1998-06-11	0.600	0.540	0.335
1998-09-11	0.419	0.429	0.301
1998-12-11	0.667	0.547	0.341
1999-03-11	0.509	0.531	0.368
1999-06-11	0.525	0.507	0.304
1999-09-11	0.419	0.449	0.343
1999-12-11	0.491	0.498	0.337
2000-03-11	0.604	0.530	0.345
2000-06-11	0.397	0.453	0.329
2000-09-11	0.466	0.450	0.369
2000-12-11	0.644	0.536	0.354
2001-03-11	0.594	0.566	0.391
2001-06-11	0.568	0.599	0.441

**Note: Quarters reported using a "floor date", e.g. any date including or after 1996-12-11 & prior to 1997-03-11 is classified as belonging to quarter "1996-12-11"*

Tables 6 and 7 display the weights assigned by the synthetic control method for the inexperienced and experienced synthetic units respectively. For the inexperienced group, minimizing the difference between the pre-9/11 Muslim treatment group and the synthetic control (Equation 2) results in the assignment of weights significantly different from zero to 11 of the 56 countries. The synthetic control for the inexperienced group is therefore a weighted average of the average grant-rates of the following countries: Algeria, Armenia, Burma (Myanmar), Ecuador, El Salvador, Ivory Coast, Jamaica, Peru, Poland, Togo, and Serbia/Montenegro.[115] The weight assigned to each of these countries conveys the proportion of

[115] The database includes several outdated country names, including Burma which is now officially Myanmar, and Serbia/Montenegro, which are now separate countries.

the synthetic control unit which is representative of a given country's data. For example, the inexperienced synthetic control for any given quarter is a function of the grant rate data from these 11 countries as follows: 0.201(Algeria) + 0.023(Armenia) + 0.370(Burma) + 0.022(Ecuador) +…. +0.098(Serbia/Montenegro).

Country Weights in Synthetic Muslim: Inexperienced *(Table 6)*

Weight	Country	Weight	Country	Weight	Country
0.201	**Algeria**	0	Ethiopia	**0.035**	**Peru**
0	Albania	0	Fiji	0	Pakistan
0.023	**Armenia**	0	Gambia	**0.049**	**Poland**
0	Azerbaijan	0	Ghana	0	Romania
0	Bangladesh	0	Georgia	0	Philippines
0.370	**Burma**	0	Guatemala	0	Russia
0	Brazil	0	Guinea	0	Senegal
0	Bulguria	0	Haiti	0	Sierra Leone
0	Burundi	0	Honduras	0	Somalia
0	Sri Lanka	0	India	0	Stateless
0	Congo	**0.068**	**Ivory Coast**	0	Sudan
0	Zaire	**0.057**	**Jamaica**	**0.074**	**Togo**
0	China	0	Laos	0	Turkey
0	Cameroon	0	Liberia	0	Ukraine
0	Colombia	0	Macedonia	0	Uganda
0	Cuba	0	Mauritania	0	Vietnam
0.022	**Ecuador**	0	Mexico	0	Yugoslavia
0	Egypt	0	Nigeria	**0.098**	**Serbia/Montenegro**
0.003	**El Salvador**	0	Nicaragua		

As seen in Table 7, for the experienced group, weights significantly different from zero were attributed to 5 of 32 possible controls. The synthetic control for this group is thus a weighted average of the average grant-rates of Albania, China, Cameroon, Russia, and Sudan as following: 0.112(Albania) + 0.642(China) + 0.068(Cameroon) + 0.173(Russia) + 0.005(Sudan).

Country Weights in Synthetic Muslim: Experienced *(Table 7)*

Weight	Country	Weight	Country
0.112	**Albania**	0	Laos
0	Armenia	0	Liberia
0	Bangladesh	0	Mauritania
0	Bulgaria	0	Mexico
0.642	**China**	0	Nigeria
0.068	**Cameroon**	0	Nicuragua
0	Colombia	0	Peru
0	Cuba	0	Pakistan
0	Egypt	0	Philippines
0	El Salvador	**0.173**	**Russia**
0	Ethiopia	0	Sierra Leone
0	Ghana	0	Somalia
0	Guatemala	**0.005**	**Sudan**
0	Haiti	0	Ukraine
0	Honduras	0	Vietnam
0	India	0	Yugoslavia

Figures 5 and 6 display the “synthetic Muslim” unit compared to the “treated Muslim” unit for the inexperienced and experienced group respectively, from December 1996 to December 2002. As indicated in Tables 4 and 5, both of the synthetic controls track the treatment units much more closely in the pre-9/11 period than does an average of all controls, which was plotted against the treatment units in Figures 2 and 3. The superior fit of the inexperienced group’s synthetic control can be largely attributed to a significantly bigger control pool (more countries to choose from), as well as to the fact that there happens to be more data used to calculate each of its quarterly averages, which reduces overall noisiness.[116]

[116] In using the term noisiness, I am not referring to the general concept of “noise” which is commonly invoked in this paper in contexts surrounding cognitive biases. In this case I am referring to the tendency of data to exhibit statistics that are far from true population parameters, the likelihood of which is increased in small sample sizes.

Figure 5: Inexperienced Synthetic Control vs Treatment

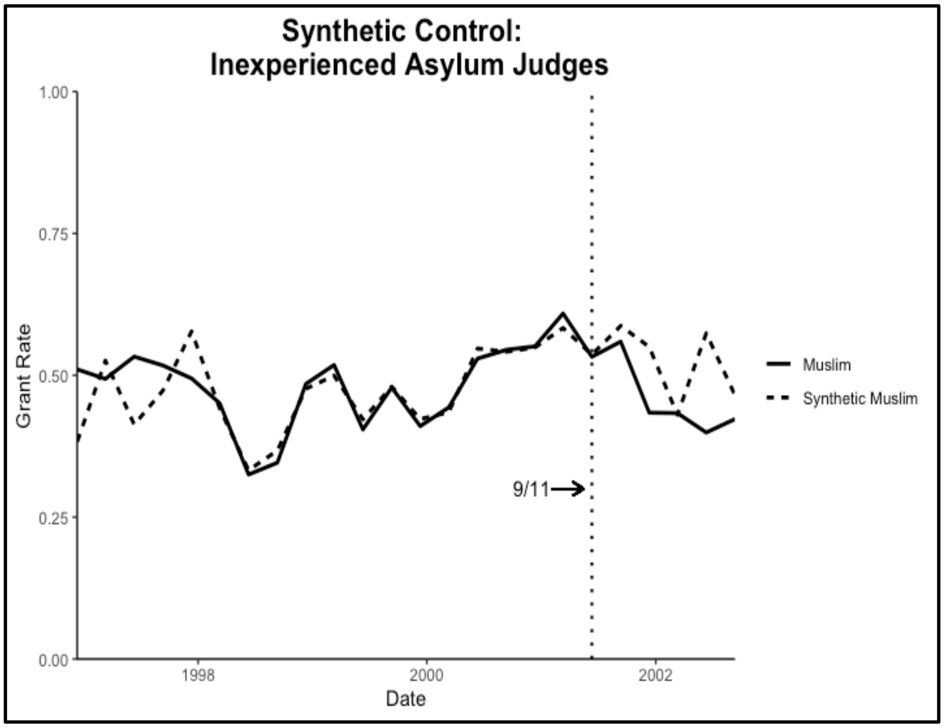
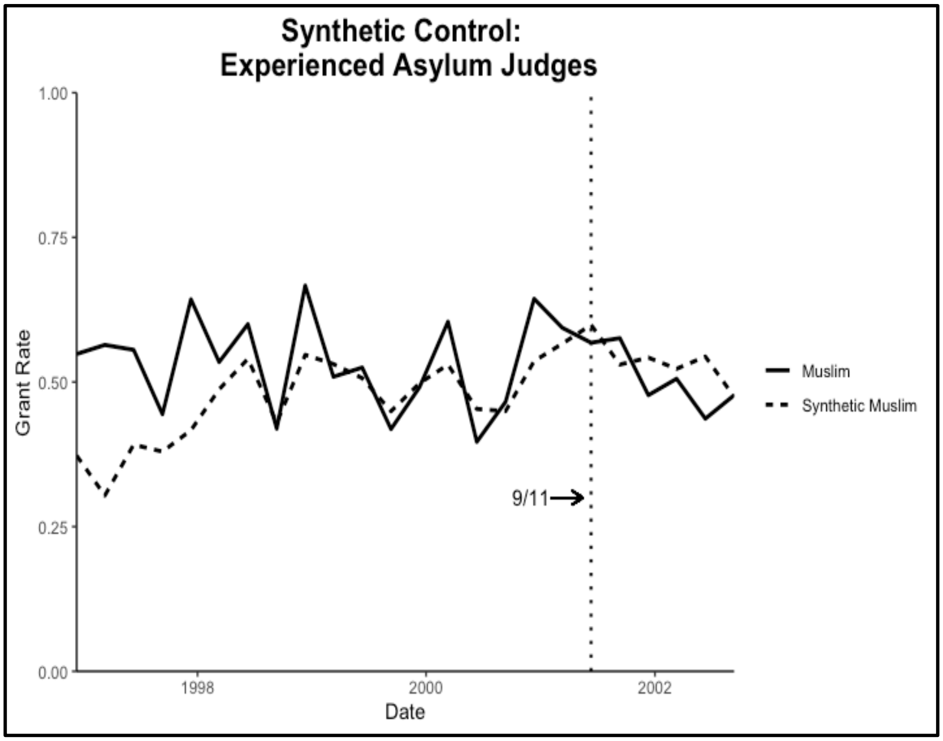


Figure 6: Experienced Synthetic Control vs Treatment



In Figure 5, the inexperienced group's synthetic control indicates that 9/11 had a negative effect on the average Muslim treatment unit grant rate. While a naïve analysis of trends in the Muslim group's grant-rate would have indicated an increase in grant-rates immediately following 9/11, it is clear from the synthetic Muslim's post-period behavior that the quarter after 9/11 indeed indicates that the attacks resulted in a drop in grant rates. This is because the increase in grant rates observed following the attacks was not as high as the counterfactual outlined by the synthetic control, which represents the trend in grant rates for the Muslim treatment group that would have occurred in the absence of 9/11.

In the post 9/11 period, the inexperienced synthetic control unit's grant rates appear to largely echo the volatility of the pre-9/11 period, while the observed treatment group's grant rates undergo significant declines, or stagnations from quarter to quarter. The Muslim grant rate experiences a small quarterly increase in the very last post-treatment period, but remains well below the counterfactual represented by the synthetic control. The period-specific treatment effects, representing the quarterly gaps in grant rate between the treatment and synthetic control, are as follows: -2.8%, -11.6%, 0%, -17.8%, -4.2%. This translates to an average post-9/11 treatment effect, across all five quarters, of a *7.1% drop* in grant rates.

Turning to Figure 6, while the experienced synthetic control fit was not nearly as clean as that of the inexperienced, it should be noted that the synthetic comparison unit succeeded in following the general trends of the Muslim treatment group fairly closely – it simply failed to match the height of the treatment group's spikes in the pre-9/11 period. It should also be noted that the relatively poorer fit does not preclude the method from generating valid treatment effects worthy of consideration and analysis. Achieving a pre-treatment fit of the quality of that in the inexperienced group is somewhat exceptional (especially for volatile data), while the experienced group's synthetic unit fit quality is fairly common.[117]

Importantly, the post 9/11 gap between the synthetic and treatment unit in this experienced group does not seem to widen very noticeably from its pre-9/11 gaps, as one would expect to observe in the presence of a significant treatment effect. Like in the inexperienced group, the Muslim treatment group experiences a net decrease following 9/11, although this decrease is less severe in magnitude. The period specific treatment effects, representing the quarterly gaps in grant rate between the treatment and synthetic control, are as follows: +4.5%, -6.5%, -1.7%, -10.7%, 0%. Overall, this translates to an average post-9/11 treatment effect across all five quarters of a 2.9% drop in grant-rates.

In order to assess the significance of the inexperienced and experience synthetic control estimates, I consider whether these results could have been a result of mere chance. How likely is it that I would have obtained results of the magnitude of these estimates if any country at random had been chosen, instead of those countries which, aggregated, form the Muslim treatment unit?

By applying the synthetic control method to each control country as if it were the treatment unit of interest, and taking the post-9/11 gaps between each control unit and its synthetic counterpart, I can create placebo treatment effects, produced under the null hypothesis of no treatment effect. If these placebo treatment effects result in gaps with magnitudes similar to those estimated for the true Muslim treatment group in either synthetic control, then it can be concluded that the original analysis fails to offer significant evidence of a negative effect of 9/11 on the grant rates of the Muslim treatment group. Otherwise, if the actual treatment effects

[117] Jones and Marinescu, "The Labor Market Impacts of Universal and Permanent Cash Transfers"; Billmeier and Nannicini, "Assessing Economic Liberalization Episodes"; Rydberg et al., "A Quasi-Experimental Synthetic Control Evaluation of a Place-Based Police-Directed Patrol Intervention on Violent Crime."

estimated for the Muslim treatment group are abnormally negative compared to the distribution of placebo effects, the analysis can be interpreted as offering significant evidence of a negative effect of 9/11 on the grant rates of the Muslim treatment group.

Following this logic, Figures 7 and 8 were produced by iteratively applying the synthetic control method used to estimate the effect of 9/11 on the Muslim treatment group's grant rate to every other country in the control pool. In other words, I proceed as if I expect the 9/11 attacks to trigger discrimination or backlash toward each of the countries in the control group, and compute the associated estimated treatment effect by taking the gap between the control unit and its corresponding synthetic counterpart in the five periods following 9/11. After this procedure is iterated over each control unit, a distribution is produced of estimated gaps for the applicants from countries that should *not* have been the target of anti-Muslim discrimination following 9/11.

Figure 7: Inexperienced Gap Plot- All Controls

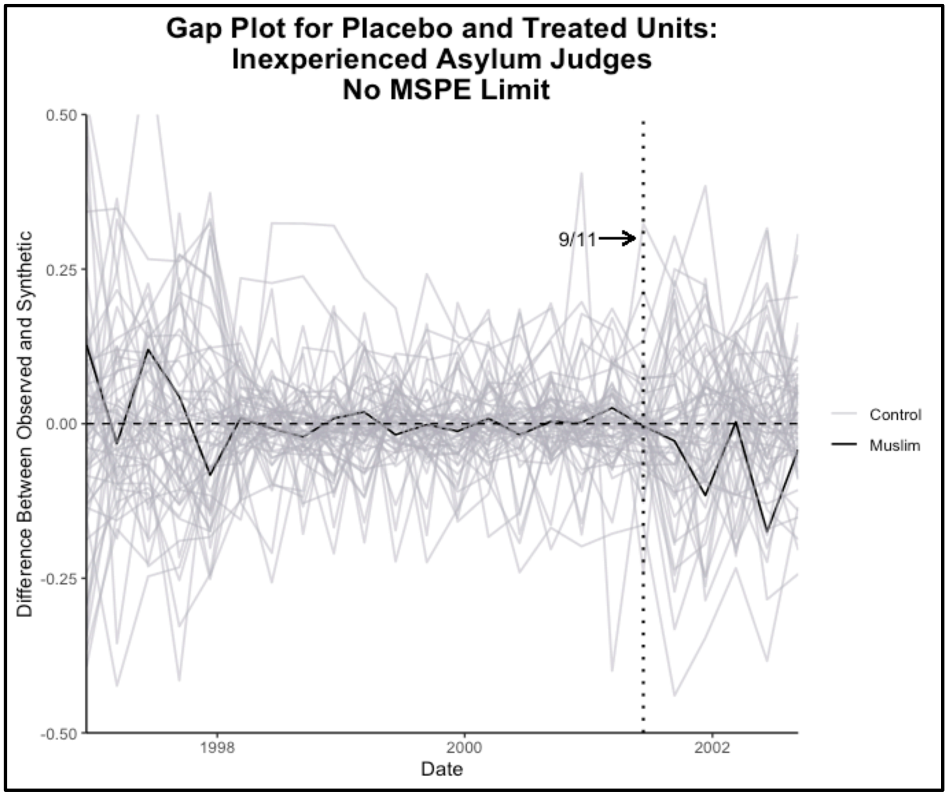
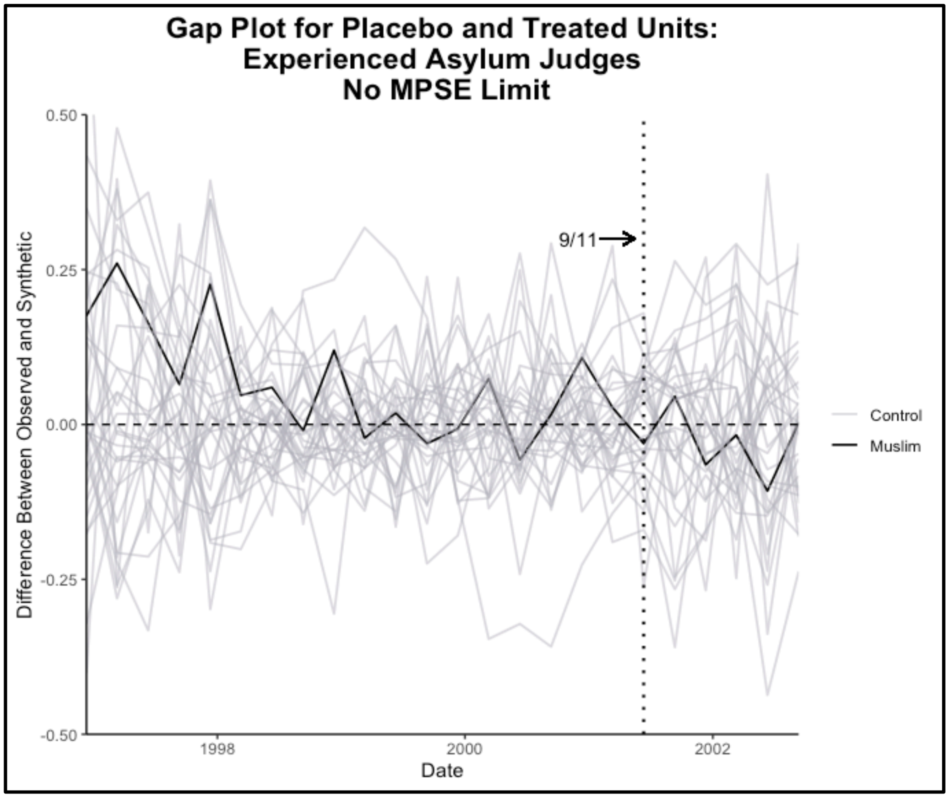


Figure 8: Experienced Gap Plot - All Controls



The grey lines in each plot represent placebo gaps, produced by each control unit's turn as faux treatment unit, and calculated by subtracting the difference between the true control unit and its synthetic unit. The black line in each plot represents the gaps observed between the true Muslim treatment group and its associated synthetic control. Neither graph appears to provide much evidence for a significant treatment effect, since both black lines are well within the range of placebo effects.

However, the size of the placebo gaps in the pretreatment period bears great relevance to the analysis of placebo distributions. Clearly, there are many control units in each experience cohort which suffer from a very poor fit in this pre-9/11 period, as indicated by the magnitude of many of the gaps. If a synthetic control exhibits a very poor fit throughout the pre-period, it is more likely to generate spurious treatment effects, by virtue of gaps that exist despite any treatment. Therefore, it is common to restrict placebo units used to create a distribution of treatment effects under the null hypothesis to units whose pre-treatment MSPE is a certain multiple of that of the treatment unit's pre-treatment MSPE. This can be thought of as a quality control meant to ensure that the placebo distribution is comprised only of gaps produced by synthetic units which closely approximated their corresponding control unit prior to 9/11. After all, it is not very useful to compare an estimated treatment effect generated by a quality synthetic control to a placebo effect stemming from a severely mismatch synthetic unit. In many cases, including these units in the placebo distribution can be misleading for inference.

In Figures 9 and 10, the placebo gaps for both the experience and inexperienced group are limited to those whose MSPE in the pretreatment period was no larger than five times that of the treatment.

Figure 9: Inexperienced Gap Plot-Controls with MSPE limit of 5x

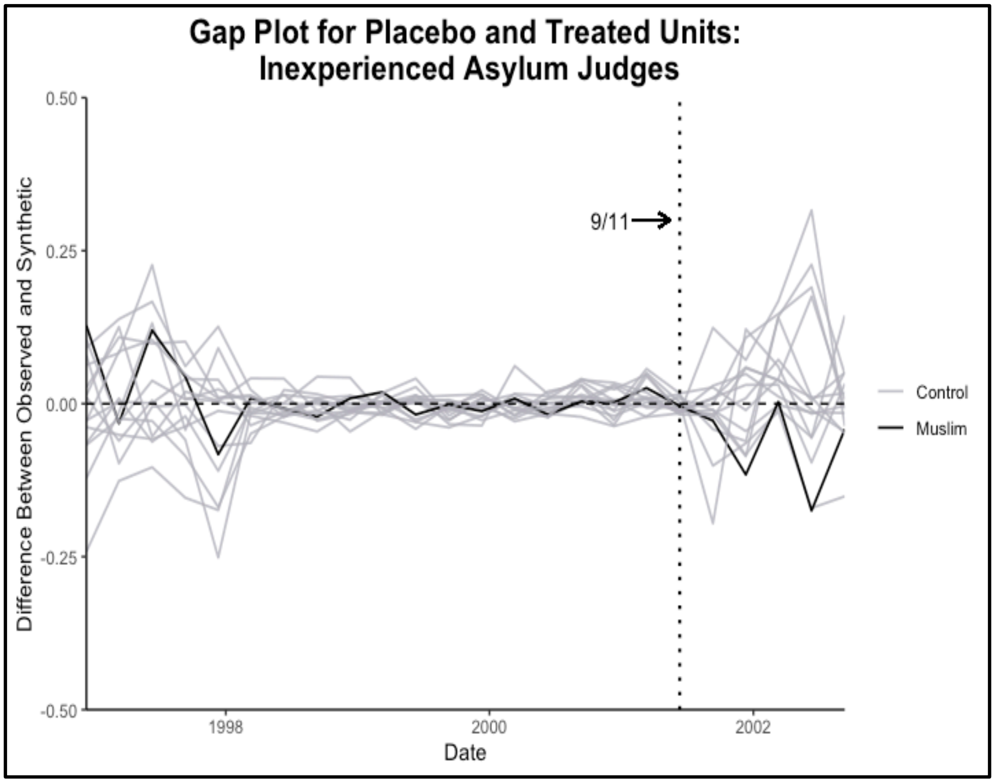
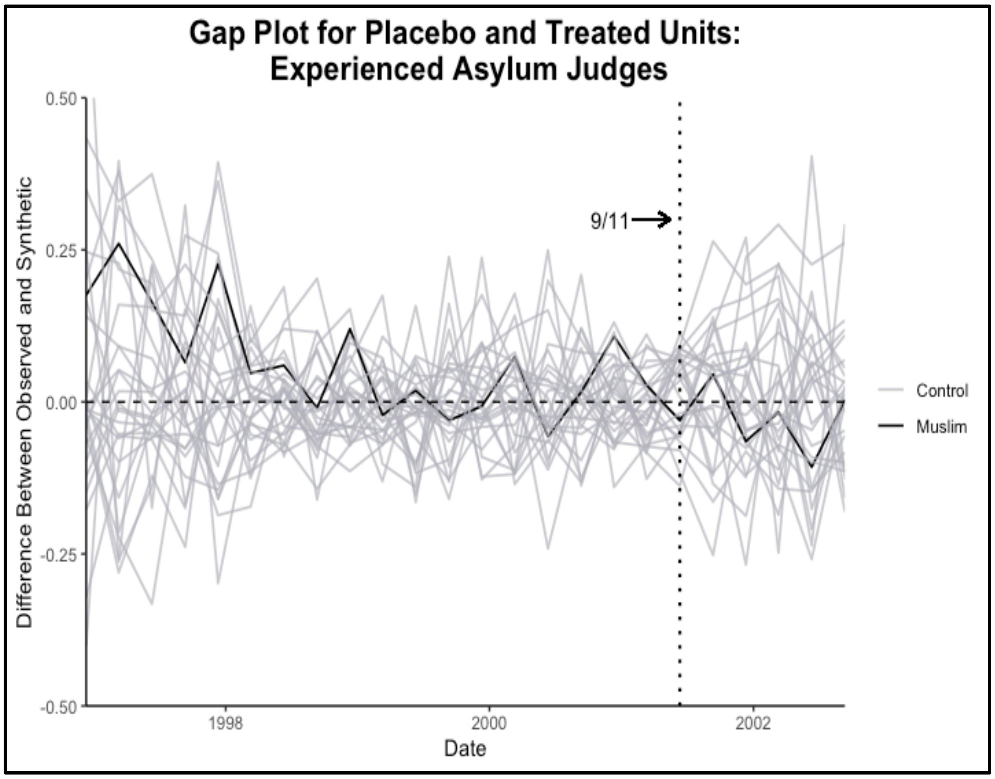


Figure 10: Experienced Gap Plot- Controls with MSPE limit of 5x



Once the countries whose synthetic control had a significantly poorer fit than the treatment's synthetic control have been discarded, the inexperienced gap plot in Figure 9 becomes much more indicative of a significant negative effect. In fact, only one other placebo unit (Cuba) for the inexperienced group generates an average treatment effect negative as the Muslim treatment group's -7.1%. In the second and fourth post-treatment period, the estimates for the Muslim treatment group (-11.6% and -17.8% respectively) are more negative than all placebos in the sample. This serves as strong, but (on its own) inconclusive evidence that inexperienced judges decreased their average grants rates for Middle Eastern Arab applicants following 9/11.

In contrast, Figure 10, the experienced cohort's gap plot, provides little evidence of a significant effect for the Muslim treatment group. The black line representing the gaps between this treatment group and its synthetic unit is generally situated in the center of the placebo distribution in the post-9/11 period. Indeed, 15 of 32 control units generate an average treatment effect more negative than the observed -2.8%. The observed treatment effects are determined to be far from abnormal under the null hypothesis of no effect. Therefore, this null hypothesis clearly cannot be rejected. Some readers may be skeptical of this conclusion given that the experienced synthetic control's pre-9/11 fit was significantly worse than that of the inexperienced. I emphasize that a poor fit is likely, if anything, to bias my results toward treatment effect estimates that are greater in magnitude, and more abnormal in comparison to the placebo distribution, not smaller estimates that fit squarely within this placebo distribution.

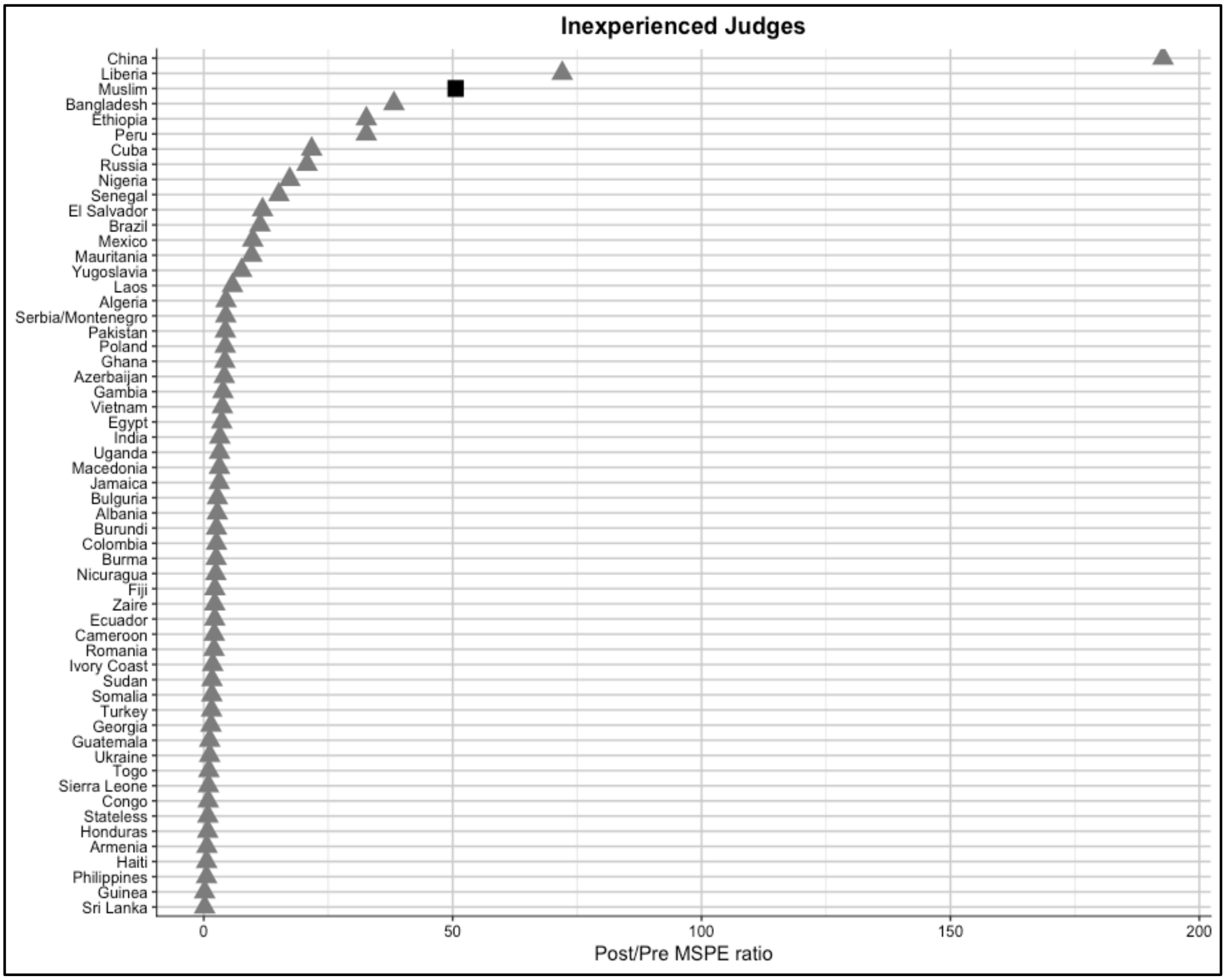
An additional method of evaluating the significance of the estimated treatment effects involves rank ordering the post/pre MSPE ratios discussed in the model section. The true treatment ranking top, or very near the top, of this list serves as evidence of a significant treatment effect. Figures 11 and 12 display the inexperienced and experienced plots respectively in which post/pre MSPE ratios are ranked. For both plots, all control countries are represented by a grey triangle, while the Muslim treatment unit is represented by a black square. Because the post/pre MSPE ratios reflect the quality of a synthetic control's pre 9/11 fit by design, I do not omit any control countries from the plots in Figures 11 and 12.

In the appendix, I display gap plots of placebo distributions paired with post-pre MSPE ratio plots based on a range of different MSPE restrictions for each experience group (Figures 14 – 21).[118] These provide an illustration of the placebo distribution associated with MSPE limits of greater than 5, although it is generally agreed that such constraints are far too lax, meaning that

[118] There is no set of figures for the experienced group with a MSPE ratio of 20 times that of the treatment unit. This is because no countries' MSPE exceeded this ratio. Therefore, the plot technically would have looked the same as Figure 8, which excludes no control countries.

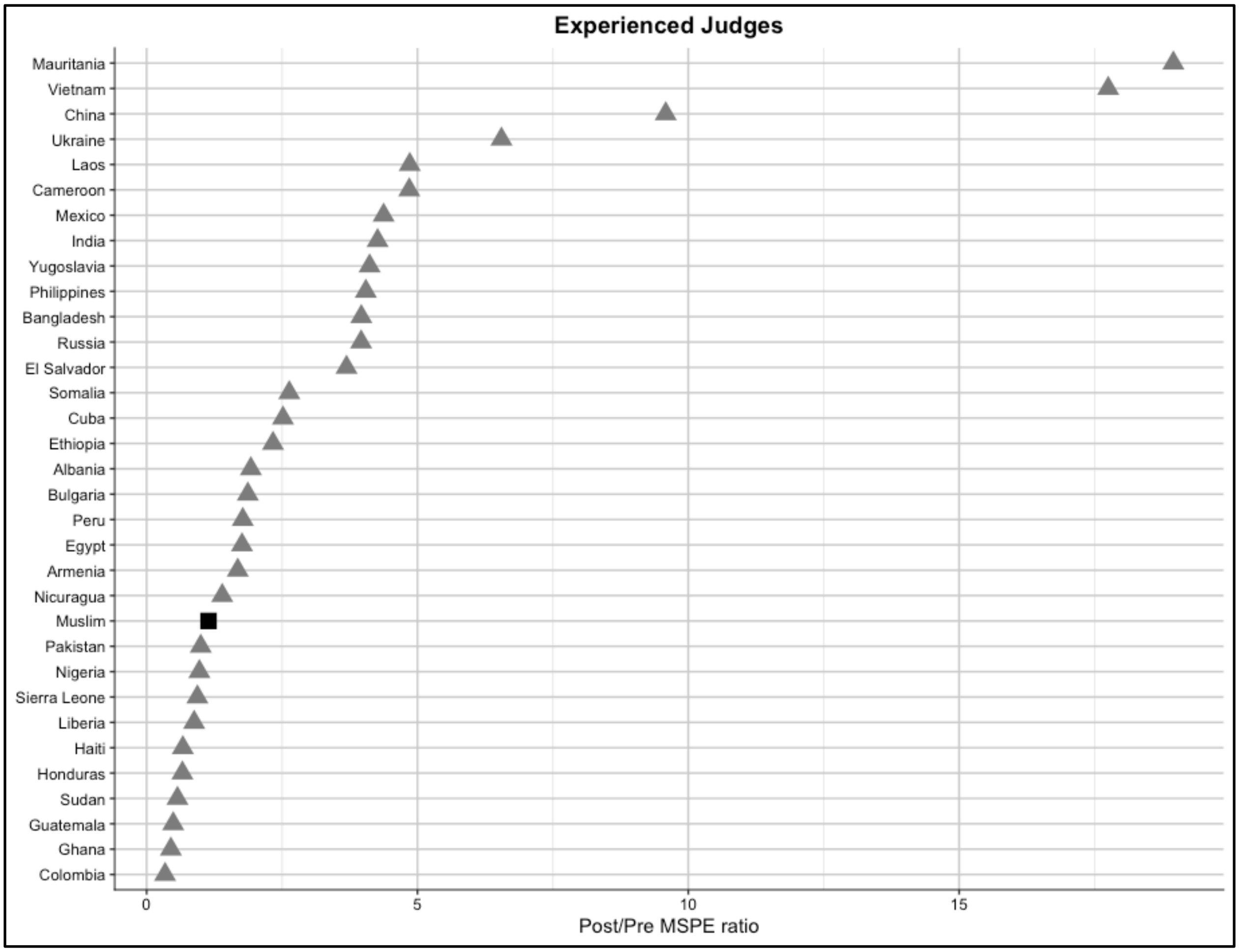
they do not impose a high enough standard of fit to warrant serious consideration.[119] To demonstrate the robustness of the experienced synthetic control's near-median position in the placebo distribution, I also include a gap plot and post-pre MSPE ratio plot with an MSPE limit of 2, so that the only countries included are those with less than two times the amount of pre-9/11 error than the actual experienced treatment unit.

Figure 11: Inexperienced MSPE Ratio Plot



[119] Abadie, Diamond, and Hainmueller, "Synthetic Control Methods for Comparative Case Studies."

Figure 12: Experienced MSPE Ratio Plot



From Figure 11, it is clear that the post/pre MSPE ratio of the Muslim treatment group is very high compared to the values for the rest of the control units. It ranks third from the top, with only 2/56, or 3.57% of control observations demonstrating larger ratios. In Figure 12, it is also clear that the post/pre MSPE ratio of the Muslim treatment group is nowhere near the top of the distribution. Rather, 22/32, or 68.75% of control observations demonstrate larger ratios. These findings serve as further evidence that among inexperienced judges, there was a significant decrease in Muslim grant rates post 9/11, while there was no significant effect among experienced judges.

Some readers may notice that China, which has a hefty influence on the synthetic control unit of the experienced group, also has one of the highest post/pre MSPE ratios of controls in this group (as shown by figure 12). Indeed, the quarterly placebo treatment effects for China are quite large in magnitude: +11.6%, +14.1%, +16.9%, +8.9%, -1.6%. It would seem that grant rate behavior for Chinese applicants changed significantly around the same time that it did for Middle Eastern Arab applicants. What implications might this have for my experienced synthetic control, which should theoretically undergo no shock or treatment, but is also 64.2% data from Chinese applicants?[120]

One can imagine that because Chinese applicants experienced such an increase in grant rates throughout the post-9/11 period, the synthetic control reflects a higher post-period grant rate than would be the case if a country with a lower post/pre MSPE had been selected in its place for the synthetic unit. Therefore, acknowledging the post-9/11 changes in Chinese data, which somewhat prevent the synthetic unit from exhibiting true "control" behavior, results in the conclusion that the experienced synthetic control was likely biased toward higher grant rates in the post-9/11 period. Critically, this does little to change the interpretation of my findings regarding the lack of evidence of discrimination for experienced judges, because a synthetic control with lower grant rates could *only* result in smaller estimates of grant rate decreases.

While some research endeavors leveraging the synthetic control method calculate a value akin to a p-value for their estimates, I do not find such a calculation to be very informative in my case. For the experienced group, it is now clear that the p-value should be very high, since all indicators point toward the fact that the treatment effect estimate is nowhere near abnormal compared to placebo units. For the inexperienced group, once the placebo distribution is

[120] The 64.2% figure comes from Table 7, in which weights are assigned to control countries by the synthetic control.

restricted by the pretreatment MSPE of no greater than five times than that of the original treatment's synthetic control model, there are only 14 other viable control units. The typical p-value calculation in the synthetic control context involves dividing the number of placebo gap estimates as or more extreme than the estimated treatment effect by the total number of control units. In my case, then, a one-tailed hypothesis that 9/11 was responsible for a decrease in the Muslim treatment group's grant rate would yield a p-value of 0.071, because one of the fourteen placebo estimates was more negative than my own estimate.

First, this p-value is problematic because of its small denominator. It is difficult to make precise estimations of p-values from a small sample size partially because, for example, a sample size of 14 mandates that all possible p-values are separated by intervals of 0.071, or 7.1% probability. The p-value distribution is far too discrete to enable inference which relies on the continuous probability of test-statistics. As Abadie, one of the researchers responsible for developing and formalizing the synthetic control method, notes in a 2019 paper, the permutation method used to generate placebo treatment effects does not attempt to approximate the sampling distributions of test statistics.[121] The p-values that can be generated from the placebo treatment analysis therefore do not carry the same meaning or significance as those generated in more familiar regression settings.

Next, how should one reconcile this "placebo p-value" of 0.071 with the p-value that could also be calculated from the pre-post MSPE ratio procedure, in which only 3.57% of control observations generate treatment effects more extreme than the treatment unit? Should I consider my p-value to be 0.0357, or 0.071? Notably, these values stratify the traditional threshold of 0.05, so that the former would lead me to deem my results "statistically significant", and the

[121] Abadie, "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects."

latter would not. It isn't clear how to precisely quantify the significance of the result generated by the inexperienced cohort. It is, however, clear that for inexperienced judges, there exists notable evidence of a negative treatment effect of 9/11 on Middle Eastern Arab asylum applicants. It is also clear that there is nowhere near as much evidence that such an effect occurred among applicants whose cases were adjudicated by judges with more than ten years of tenure under their belt. These findings reinforce the theory that greater susceptibility to cognitive biases among inexperienced judges renders these judges more likely to rely on stereotypes, and exhibit greater racial and cultural biases.

Discussion

Overall, my findings from both phases of analysis are consistent with the theory that the cognitive immaturity of less experienced judges results in higher discrimination based on racial and religious stereotypes. The results of my learning curve analysis were consistent with that of Chen et al. (2016) in indicating that as judges accrue experience, they exhibit significantly less evidence of the gambler's fallacy.[122] In other words, more experienced judges can be understood to base their decisions more upon the features of an asylum hearing, and less upon irrelevant factors.

Beyond confirming the findings of Chen et al. (2016), my learning curve analysis proves novel in its demonstration of the significance of the ten-year benchmark, as well as its demonstration of the overall shape of the curve. From the learning curve model, it is apparent that asylum judges exhibit the greatest amount of cognitive bias in their early years of tenure, and can generally expect a yearly increase in decision quality until they have completed their

[122] Chen, Moskowitz, and Shue, "Decision Making Under the Gambler's Fallacy."

tenth year. After this point, they do not see further improvement. It is therefore evident from my research that the relationship between cognitive biases and experience is characterized by diminishing marginal returns, as evidenced by this eventual plateau in yearly improvement.

With respect to the limitations of the learning curve analysis, it is unclear how generalizable the gambler's fallacy findings are to other cognitive biases. It seems possible that different cognitive biases might exhibit different learning curves, as well as different points at which improvements are diminished to zero. However, because so few cognitive biases are as suitable for empirical identification as the gambler's fallacy, the curve measured in my paper is the best available proxy for general maturity with respect to cognitive biases for the time being.

The results of my 9/11 discrimination analysis are consistent with Arnold et al. (2018) in showing that inexperienced judges exhibit more discriminatory behavior than their experienced counterparts.[123] My results are also consistent with Mensah and Opoku-Agyemang (2018) in detecting backlash against Middle Eastern Arab respondents in immigration court following a terrorist attack, although my research suggests that this backlash is largely driven by inexperienced judges.[124] My findings are novel in an empirical sense, and in a broader theoretical sense as well. Empirically, I conduct the first-known analysis on the effects of 9/11-related anti-Muslim sentiment on asylum case outcomes, providing evidence that the attacks significantly decreased the grant rate for applicants from Middle Eastern Arab countries. Critically, this decline in grant rates is only significant among inexperienced judges, who, by virtue of having less tenure than the ten-year benchmark, are also documented as more prone to cognitive biases.

[123] Arnold, Dobbie, and Yang, "Racial Bias in Bail Decisions."

[124] Mensah and Opoku-Agyemang, "Innocent Until Stereotyped Guilty? Terrorism and US Immigration Court Decisions."

In sorting the experience groups according to the ten-year milestone, I deliberately construct an experience variable that corresponds directly with measurements of cognitive bias. By showing that this specific inexperienced group of judges exhibited behavior consistent with cultural discrimination, I am able to do what neither Chen et al. (2016) nor Arnold et al. (2018) can, taking the cognitive immaturity associated with inexperience, and linking it very directly to discriminatory behavior. My findings are therefore theoretically novel in being the first to provide such strong evidence for the mechanism by which susceptibility to cognitive biases due to inexperience results in racial, ethnic, and/or religious discrimination.

Ultimately, while the 9/11 discrimination evidence is suggestive of the mechanism proposed above, it is not fully conclusive. An alternate explanation that would also be consistent with my findings is that judges simply become more rigid and fixed in their decision patterns as they gain tenure. In this model of judicial behavior, less experienced judges can be thought of as more open to new information as they seek for themselves an eventual stable consensus regarding how to treat cases of a particular character or group. An exogenous shock such as 9/11 would therefore be more likely to cause an adjustment in the behavior of inexperienced judges, while having little effect on experienced judges who have already "made up their mind" about the credibility of the average claim of a Middle Eastern Arab applicant.

In some cases, decision stability is desirable, as something like 9/11 should have no bearing on the merits of asylum claims from these applicants (barring any subsequent changes in case quality composition). In a scenario like the outbreak of the Syrian civil war, however, more flexibility would be desirable, so that judges account for the increased threat faced by Syrian applicants, and increase grant rates accordingly. From my data alone, it is impossible to distinguish between this explanation and my proposed mechanism. However, when incorporating the nature of the findings of Arnold et al. (2018), which argues that inexperienced judges demonstrate more racial bias consistently, and in the absence of an exogenous shock, the alternative explanation of stability via experience seems less likely to have driven my 9/11 results.[125]

Policy Implications & Recommendations

The analysis conducted in this research endeavor has revealed evidence, not about the effectiveness of a certain policy or program, but about the evolution of judicial decision-making, and the implications of that evolution. As a result, making policy recommendations is somewhat less straightforward than might be the case in program evaluation-oriented research. Even if one heeds the claim presented in this paper: that less experienced judges are more susceptible to cognitive errors in decision-making, and that these errors can give rise to patterns of discrimination across racial or religious lines, it isn't obvious which policies are best poised to address such seemingly intractable flaws in the justice system. In this section, I contemplate the feasibility and attractiveness of potential policy solutions, and conclude by issuing final recommendations.

One of the main takeaways from this research is of course that inexperienced judges impose a significant cost to those applying for asylum, as well as to the integrity of the justice system. It would be nonsensical, however, to respond to this finding by barring inexperienced judges from rendering decisions on asylum applications. This would merely preclude them from gaining the very experience that appears to convey elevation to a more desirable level of decision maturity and stability. It might be reasonable, though, to institute an apprenticeship period in

[125] Arnold, Dobbie, and Yang, "Racial Bias in Bail Decisions."

which an inexperienced judge shadows a more experienced judge, and observes as she presides over cases and make decisions on applications for relief. Using insights from the learning curve traced out in this paper regarding the amount of tenure that the average judge must accrue before achieving their peak and final level of decision maturity, the experienced judge in such a partnership would ideally have at least ten years of experience in the immigration courts.

This apprenticeship approach would yield desirable results under the assumption that the actual act of managing and executing a decision is not what specifically improves decision-making skills. Rather, it relies on the benefits of mere exposure to, and consideration of, factors integral to the decision-making process, and treats the benefits acquired from this passive capacity as a reasonable, albeit imperfect, proxy for experience. Further research could investigate how this type of sidelined role might be conducive to accelerating the decision maturity process.

The ideal length of such an apprenticeship period is far from clear, as there are likely diminishing marginal returns to the decision maturity gained from this position. The apprenticeship would also impose a cost on the court system in the sense that it is effectively reducing the number of active judges, straining a system that is already overburdened with a backlog of over half a million cases. To avoid further crippling the immigration courts and exacerbating the current shortage of judges, the apprenticeship period would ideally occur during a judge's very first stage of service, and only last as long as research determines is necessary to improve resilience to cognitive biases significantly from baseline. Even if the program cannot close the entire gap between inexperienced and experienced decision maturity, the stakes are high enough in the asylum context that marginal improvements will give thousands more asylum seekers per year the more merit-based consideration their (often life-or-death) cases deserve.

Furthermore, the process by which novice judges are assigned to experienced counterparts deserves careful consideration. As discussed previously, there are enormous discrepancies between immigration judges in their asylum decision patterns, especially along the lines of applicants' country of origin. It is possible that the idiosyncratic decision tendencies of the experienced judges will impact the takeaways gleaned by the inexperienced judge observing them. Additionally, if the "apprentice" judge disagrees with the reasoning of or the decisions made by their mentor, they may discard the value of learning from the more experienced judge's behavior, become less engaged, and reap fewer benefits as a result.

A possible positive side effect of an apprenticeship policy could be the increased oversight and accountability for the experienced judge, which comes with the knowledge that someone is intently watching their decisions under the pretense of a learning environment. Given the very little oversight, and large amount of discretion currently experienced by immigration judges, the introduction of a sharply observant presence may force experienced judges to think more carefully about the soundness of their justifications, thereby further reducing room for the influence of cognitive biases. It is also possible that pressure may cause the rulings on asylum decisions of experienced judges to trend more towards normative expectations (what judges thinks they are *supposed* to do), rather than following their own instinct or lines of reasoning. It isn't clear whether an adaptation to better meet normative expectations would exert a positive or negative consequence on the accuracy of a judge's decision.

Because of the state of the backlog, a preferable alternative to the apprenticeship program might be a simulation tool which is integrated as a mandatory component of training for new judges before they are allowed to rule on asylum applications. The tool could be used to present components of cases that have been presented before real immigration judges in the past.

Additionally, the selection of cases could be cherrypicked to ensure a wide variety of exposure to asylum applicants of different nationalities, defensive statuses, or legal representations. This would provide a distribution of cases distinct from that experienced in many immigration courts, seeing as different courts tend to receive starkly different compositions of nationalities, and other case attributes. In the spirit that diversifying exposure should broaden experience in a way that multiple cases of similar attributes might not, a simulation case, on the margin, has the potential to be more beneficial to a judge's decision maturity than would the average case adjudicated at the judge's court of employment. Similarly to the apprenticeship policy, further research would have to be conducted in order to determine the structure of diminishing marginal returns associated with exposure to these simulated cases, and ultimately the amount of training via this simulation tool that is appropriate and efficient.

The usefulness of the simulation tool could be increased if the compilation of cases was purposeful with respect to the "true outcome" of each case. By true outcome, I refer not to the judge's final decision on the asylum application, but to what happened to the applicant in the short-term (two years or so) following the judicial decision. In most cases in which asylum was granted, it is difficult to observe the applicants in the period after this decision and confirm whether that person truly qualified for asylum by law. However, if after asylum was denied and an applicant was deported back to their country of origin, that applicant fell victim to one of the forms of persecution specifically delineated in asylum law, it seems clear that the judicial decision made was not just wrong, but wrong at a very high cost. Therefore, the tracking of applicants who are denied and subsequently deported could provide very useful in understanding the ways in, and extent to, which immigration judges make these grave errors.

Until recently, such tracking was nothing more than a hypothetical. Recent human rights advocacy groups, however, have recognized the value of documenting the aftermath of denied asylum claims. As mentioned previously, a February 2019 report by Human Rights Watch, identified 138 cases of Salvadorians who had been killed since 2013 following their deportation from the United States.[126] As disturbing as this data may be, more like it should be leveraged to improve the asylum deliberation process, and make clear its harsh consequences. If the simulation of each case is followed by evidence of the actual applicant's true outcome, the judges may become more skilled at diagnosing whether cases that do or do not meet the criteria for asylum. Such feedback would constitute a major improvement over simply regarding the outcome of each case to be the actual decision handed down by the judge, because in many ways, the judge's decisions is merely her best attempt to predict the case's true outcome. The true outcome is actually representative of the material consequences of the decision.

Just because this analysis indicates that more experience translates to a smaller influence of cognitive biases does not mean that the only solution to judicial errors is finding a way to supplement the experience of novice judges. A more direct approach could yield greater benefits at a lower cost of foregone labor to the courts. It is therefore worth exploring the administration of training that directly builds resilience to cognitive biases. The following considerations in this section will focus on how to mitigate cognitive biases more generally within the courts. This approach presents the potential benefit of increasing decision performance across all experience levels, as compared to the previously considered strategies aimed mostly at "imputing" experience onto the inexperienced judges.[127]

[126] Kennedy, Parker, and Human Rights Watch (Organization), *Deported to Danger*.

[127] Just because a strategy that directly targets cognitive biases is likely to assist more experienced judges as well does not imply that these judges will gain as much as their less experienced counterparts. Rather, it is likely that this group will see a more muted effect.

Certain primers or cues, which can seem insignificant, have the power to make people radically more or less susceptible to cognitive biases. When people are asked to slow down, and think deliberately about a problem or situation, research has found that they use an entirely different part of their cognitive system to execute a subsequent decision.[128] Critically, the decision resulting from a deliberative thinking process tends to be calibrated more by the reality of the decision-maker's situation, and less by extraneous information or irrational impulse.[129] Education promoting a kind of meta-cognition, by increasing awareness about the prevalence of cognitive errors, even in high-stakes decision making, has been proven to be an incredibly successful tool against these biases, especially in the medical field.[130]

Therefore, all new judges, before being officially deployed to the courts, could undergo a training in which they learn about such primers, and gain skills and awareness that prove useful for recognizing their own susceptibility, and for introspecting in order to perform deliberative, rather than automatic decision making on asylum applications. Intervention efforts aimed at driving a habitual cognitive transition from the automatic to the deliberative have proven incredibly effective in contexts such as crime reduction among economically disadvantaged youth, and there is no obvious reason as to why they would not yield results in a judicial context as well.[131] Indeed, this type of training has already been recommended specifically for judges, by some of the first researchers to confirm that judges were alarmingly susceptible to a range of cognitive biases.[132]

[128] Kahneman, *Thinking, Fast and Slow*.

[129] As a clarifying point, thinking "slow" doesn't translate to a significant extension of case length. Rather, I use slow in the spirit of Kahneman (2011), referencing a deliberative reasoning process rather than the automatic one. Judges thinking "slowly" most likely would involve their taking a few seconds to get into the correct headspace.

[130] Croskerry, "The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them."

[131] Heller et al., "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago."

[132] Guthrie, Rachlinski, and Wistrich, "Inside the Judicial Mind."

Importantly, there is also evidence that the effect of the gambler's fallacy can be mitigated by increasing the literal time that elapses between subsequent decisions.[133] Such mitigation is attributed to the decline in salience of the previous decision as the time elapsed increases. Again though, the immigration courts are already facing tremendous backlog, and further reducing the speed at which cases are processed is an unattractive option. However, if it is true that time is an effective mitigation mostly because of its correlation with waning salience, it is possible that intentional efforts to directly reduce the salience of a judge's previous case could yield similar promise.[134]

Additional research conducted in the medical field has found that reminding physicians of the value of Bayesian-type reasoning prompts them to draw less on their most recent patient, and more on their combined knowledge from all relevant patients they have seen throughout their career.[135] Similar tactics could be applied to judges between cases, to decrease the salience of the previous ruling, and simultaneously provide an indirect reminder of the reality that cognitive biases are likely to be at work in the courtroom. It should be noted that such a tactic, while addressing concerns about the gambler's fallacy specifically, may not be extremely effective in combatting other behavioral biases not evidenced in this paper, which are still likely to play a role in judicial decisions. This intervention strategy is therefore best incorporated within the more comprehensive cognitive bias training discussed previously.

A less direct, but seemingly promising, strategy to combat cognitive biases involves imposing the requirement that all asylum application decisions be accompanied by a formal

[133] Gold and Hester, "The Gambler's Fallacy and the Coin's Memory."

[134] Camacho, Donkers, and Stremersch, "Predictably Non-Bayesian."

[135] Nisbett et al., "The Use of Statistical Heuristics in Everyday Inductive Reasoning"; Hall, "Reviewing Intuitive Decision-Making and Uncertainty."; Bayesian reasoning involves incorporating conditional probabilities into decision-making, and updating these probabilities when new evidence is provided.

written decision produced by the judge. In this decision, the judge must give a detailed account of their reasoning process and justification. The logic accompanying this requirement is that by forcing a judge to write down their thoughts and rationale, she will necessarily engage in deliberative thinking, which reduces the chance that she make a decision on the basis of features irrelevant to the merits of cases. Also, the requirement of written decisions should serve to hold the judges more accountable, in the sense that their decisions could be requested and come under scrutiny by a superior, or the appellate court, at any time. Importantly, that written decisions necessarily meet a minimum length should not be a priority. Rather the written decision requirement could be mandated to follow a certain structure, or template that accounts for the existing duress of the backlog. This template should be designed in cooperation between the Justice Department and judges, so as to balance the competing goals of efficiency and thoughtfulness in the adjudication process. So long as the judge is forced to think about the key facets of the case via the template, there is no need for the written decision process to engulf her time and exacerbate the severity of the backlog.

It seems reasonable to relax the expectations regarding length and thoroughness on the written decisions by individual immigration judges, especially when considering the improvement a quick written template should yield over the status quo of no requirements on written decisions. For the BIA, however, higher standards should be required. Given the role that the BIA plays in adjudicating especially controversial, and the fact that it faces (necessarily) a much smaller case load than the court system at large, written decisions of conventional length and quality should be demanded. This requirement would not be novel – written decisions were mandatory for the BIA until 2002, when John Ashcroft scrapped the rule as part of a sweeping series of reforms which removed many bureaucratic checks and oversight functions from the immigration courts.[136]

Notably, strong incentives for accuracy have illustrated potential to reduce the influence not just of the gambler's fallacy, but of a variety of other frequent cognitive biases.[137] There are two major obstacles in the current immigration system, however, which prevent this finding from being effectively leveraged. First, as discussed previously, true accuracy is essentially unknown in asylum cases because of a complete lack of data. Accuracy is difficult to incentivize if it is unverifiable. Second, there is virtually no oversight on the behavior of immigration judges (evidenced in part by staggeringly different decision patterns among judges operating in the same court). Judges are largely free to decide cases as they please, and are not forced to provide solid reasoning for their decisions due to the lack of requirement of a written decision, as discussed previously. While there exists channels through which to file complaints about the conduct of specific judges, these complaints very rarely result in significant disciplinary actions against a judge in question.[138] There are practically no structurally or organizationally-driven incentives for a judge to achieve high levels of accuracy.

With regard to the first obstacle to promoting accuracy in the courts, is difficult to overstate the value of acquiring and disseminating knowledge about the true outcomes of asylum cases which result in deportation. Such data has the potential to prove incredibly helpful in impacting the calibration of judicial decisions, given that, for most, it will be their first truly accuracy-based training: the first time seeing substantial proof of an incorrect decision.[139] If the

[136] Vaala, "Bias on the Bench."

[137] Chen, Moskowitz, and Shue, "Decision Making Under the Gambler's Fallacy"; Thaler and Sunstein, *Nudge*.

[138] Rosenberg, Levinson, and McNeill, "For U.S. Asylum Seekers, Some Judges Are a Better Bet than Others."

[139] I say "for most" because I cannot affirm for sure that no active judge hasn't been contacted with information retroactively about the status of a person to whom they denied asylum.

immigration courts themselves are not willing or able to undertake the necessary responsibility of data collection, the Department of Homeland Security should be willing to fund the execution of these tasks via a contract with a human rights advocacy group such as Human Rights Watch, which has already proven its earnestness in such efforts.

To be clear, the tracking of each individual asylum applicant who is deported would be far too onerous for any one organization to take on as a project. Rather, select individuals should be randomly sampled, so as to create a diverse set of "data points" regarding accuracy. It should be clarified that total accuracy is fairly impossible to estimate due to the difficulty associated with determining the accuracy of a decision to *grant* asylum. Due to the enormously higher costs associated with an erroneous denial compared to those of an erroneous grant, I argue that an optimization of accuracy within denial decisions is the best use of resouces related to the goal of improving accuracy.

To address the second obstacle to accuracy incentivization, measures to increase accountability of judges should be enacted across the immigration courts. One such measure could involve increasing the consequences of a certain number of complaints within a year. Currently, very few punitive measures are taken against judges who receive high numbers of complaints, and those that are taken tend to be weak, often amounting to a slap on the wrist.[140] In 2016, for example, the Executive Office of Immigration Review received 138 complaints against immigration judges, which included allegations of bias and concerns about due process and judicial conduct more generally. Only 102 complaints were resolved within a year of their filing date. Only three complaints resulted in "discipline" (defined as reprimand or suspension of the

[140] Rosenberg, Levinson, and McNeill, "For U.S. Asylum Seekers, Some Judges Are a Better Bet than Others"; Misra, "DOJ Changed Hiring to Promote Restrictive Immigration Judges."

judge involved).[141] As discussed previously, an additional measure to increase accountability may be mandatory written decisions. These decisions would be available for scrutiny by the Department of Justice upon receipt of a complaint, or the occurrence of another controversy.

If data collection on deported applicants manages to improve substantially, each judge will eventually accumulate a track-record illustrating the consequences of their errors. Keeping in mind that judges are only human, it is unreasonable to expect these statistics to propel judges to achieve an error-free level of performance. These statistics should, however, provide an incentive to improve, as well as to carefully weigh the grave costs of an erroneous denial. Indeed, statistic-tracking along with dashboard reporting systems have proven very successful in promoting improvements, even with respect to cognitive biases, in a variety of professional fields.[142] If steps to make accuracy measurement feasible are taken, and greater accountability measures enacted, accuracy will be much more highly incentivized than in the current system, which should, in turn, drive better decision quality across the board.

Finally, I echo a less novel, but extremely important recommendation which has been made by immigration judges, politicians, and policy researchers alike for years – to increase the size of the BIA, and significantly increase the number of immigration judges in the courts.[143] The current backlog as it stands is unsustainable, not to mention an incredible violation of the human rights of those awaiting adjudication. This problem has not gone unnoticed by the Department of Justice, and previous attempts have been made to hire judges in bulk and ease the strain on the courts. In an eighteen month period ending in 2017, 79 new immigration judges were

[141] Rosenberg, Levinson, and McNeill, “For U.S. Asylum Seekers, Some Judges Are a Better Bet than Others.”

[142] Thaler and Sunstein, *Nudge*.

[143] Vaala, “Bias on the Bench”; Yeh, “Today’s Immigration Legal System.”

appointed.[144] During this period, the rate of new cases also climbed. As a consequence, little to no headway was made in reducing the backlog. Waiting times still hovered around an average of three or four years in many large urban courts.[145] To combat a backlog of this magnitude, it is clear that hiring needs to be scaled up dramatically.

Rather than hire judges arbitrarily when the backlog increases or becomes more pressing for whatever reasons, judges should be appointed more systematically, in a way that is directly linked to the current wait times in individual courts. Records of average wait times are currently kept for all immigration courts. A benchmark wait time deemed acceptable (or at least a reasonable short-term improvement), for example: one year, could be proposed. Judges would be hired and appointed to certain courts where the one year wait time is surpassed by a given length, whether that be two, or maybe six months. Courts would be ranked based on the severity of their violation of this average wait time. Hires and appointments would be allocated accordingly.

This type of funding structure would require approval by Congress, who is currently responsible for appropriating additional funding for immigration judge appointments. However, this new hiring framework may be a welcome change for Congress due to its precise allocation process and ties to documented need among courts. As opposed to previous funding agreements, Congress will be ensured that each judge is being hired out of necessity, and will not be placed arbitrarily within the court system, without any guarantee of actually alleviating the backlog. Additionally, doubling the size of the BIA, which is a reasonable request given the backlog that it too faces, requires the hiring of only fifteen judges. This represents a mere fraction of the hiring surges that have previously taken place for the immigration court as a whole. Among

[144] "Despite Hiring, Immigration Court Backlog and Wait Times Climb."

[145] "Despite Hiring, Immigration Court Backlog and Wait Times Climb."

many other benefits to the court system at large, a reduction in backlog should encourage judges to take their time with cases, which is especially relevant to my research findings. Increased time spent on a case should improve decision quality, as well as limit reliance on automatic thinking techniques due to a time-squeeze, or the crushing burden of a massive docket.

Ultimately, evidence from my research unfortunately does not shed light on what policies would be most beneficial in addressing the cognitive biases in the courtroom which predominantly influence, and promote discriminatory behavior among, less experienced judges. A lack of obvious policy solutions however, does not make it less costly to abstain from attempting to adjust the asylum system accordingly. My research produces evidence that the inexperience of, and biased reasoning by, immigration judges is costing thousands of asylum seekers a fair deliberation of their claims. The current immigration system thwarts humanitarian interests by ignoring this gaping flaw. Ultimately, I can recommend with confidence that the immigration courts take four immediate actions, listed in order of priority. Should significant resource constraints prevent the unilateral implementation of any of these policies across the court system, my findings make it clear that inexperienced judges should be prioritized for whatever programs are able to be enacted.

First, I recommend that the immigration courts implement mandatory cognitive bias training, mirrored off of successful programs that have been historically administered to physicians. This training should be completed upon appointment, and depending on cost of administration, periodically throughout a judge’s tenure as well. This policy seems the most direct, and probably the most cost-effective solution to the problems outlined in this research.

Second, I recommend that comprehensive efforts to understand what happens to asylum seekers who have been deported are undertaken, whether directly by the Department of

Homeland Security, or via a contract with a trusted organization. By not tracking the outcomes of these applicants, the United States is able to dodge the consequences of their asylum system, and is hopeless at gauging any real or useful form of accuracy. Additionally, the judges are being deprived of information which is likely to prove extremely helpful in calibrating their decisions, and incentivizing accuracy. This policy will be costly, but transformative to the current understanding of consequences in asylum court. Also, the cost can be modified without many repercussions, under the assumption that careful sampling techniques are undertaken by those managing data collection.

Third, I recommend that the Department of Justice impose the requirement of written decisions for all judges, albeit in a modified form discussed previously, which avoids placing significant burdens on the judges. For the Board of Immigration Appeals specifically, pre-2002 written decision requirements should be reinstated. This will increase accountability of judge decisions, as well as make mandatory a thorough introspection process, thereby forcing deliberative reasoning.

Finally, I recommend that structural changes are enacted across the immigration courts. This primarily involves the increase in immigration judges, as well as in the size of the BIA. Rather than bulk hiring sporadically, the immigration courts should be held to a certain wait time standard, and hire in accordance with breaches of that standard. Accountability standards should also be raised, meaning greater punitive measures for judges with complaints filed against them (especially those with many complaints). A mere "reprimand" is not an acceptable form of discipline against a judge who systematically and egregiously deprives applicants of a fair trial. Disciplinary procedures should be adjusted to more frequently threaten and enforce suspensions, and even dismissals of judges, in extreme cases of complaint receipts.

In addition to curtailing inappropriate behavior, a more punitive complaint system should aid in incentivizing accuracy, which typically results in better decision-making. These structural changes are important, but require coordination and mobilization across all branches of government, which renders implementation a time-consuming and significantly expensive process. The enactment of these changes should certainly yield improvements with respect to decision-making influenced by cognitive biases, but would address more directly other flaws in the system, such as the backlog, which are dire problems, but not the focus of this paper. I therefore encourage prioritization of the first three proposals due to their increased germaneness to my research, higher likelihood of time-sensitive implementation, and greater feasibility.

Because the performance of the apprenticeship or simulation tools are unknown and less supported by existing research, I am not in a position to recommend for implementation either of these policies at the moment. Research first needs to be conducted on the relative efficiency and effectiveness of each program, which requires careful randomized control trial designs and likely several years of follow-up data. However, given the high-stakes nature of the issue, this research should be commenced as quickly as possible. If either of these policies illustrates a greater effectiveness than any of the solutions in place, it should accompany, or potentially replace whichever programs are respectively lacking. The stakes of asylum cases are simply too high to allow inexperienced judicial decisions, or any decisions which are a product of cognitive biases, to proceed unfettered. If the United States justice system truly idealizes the right to a fair trial, the decision to divert resources to programs and policies which reduce the role of noise and randomness in the courtrooms should be made swiftly, and with conviction.

Conclusion

Ultimately, my research identifies the evolution by which judges become less susceptible to cognitive biases as they accrue tenure, and finds that inexperienced judges, who are less resilient to these biases than their experienced colleagues, are also more likely to engage in discriminatory behavior. It provides novel contributions to discussions both regarding the effect of experience on cognitive biases, and regarding anti-Muslim discrimination in American asylum courts following 9/11.

By tracing out a learning curve, which measures evidence of the gambler's fallacy associated with a certain number of years of tenure, I find that judges appear to become less susceptible to cognitive biases with each passing year, until they have more than ten years of experience. After this ten-year benchmark, significant gains against cognitive biases are not observed with additional years of experience. Because of this, the judicial learning curve exhibits diminishing marginal returns, as the cognitive gains from experience are greatest in a judge's initial years, and finally level out to effectively zero once a judge has served for ten years. The "height" of the learning curve is also informative. In practical terms, the height of the learning curve indicates that if a judge's previous asylum adjudication is a grant, that judge is 7.3% more likely to deny her next case if she is in her first year of tenure than if she is in say, her fifteenth, or even her eleventh year.

Armed with the knowledge that judges with more than ten years of tenure can be considered to have reached a final "cognitive maturity", I divide judges into two groups based on their tenure relative to the benchmark. I then analyze the extent to which higher levels of cognitive bias in the inexperienced group correspond with higher levels of anti-Muslim discrimination following 9/11. Using the synthetic control method, I find evidence that inexperienced judges indeed exhibit discriminatory behavior after 9/11. I find no such evidence for their more experienced counterparts. Overall, I estimate that in the fifteen month period following 9/11, grant rates for asylum applicants from Middle Eastern Arab countries dropped 7.1% among inexperienced judges. My comparative estimate for experienced judges over this period is a 2.9% drop in grant rates, but this was determined not to be significantly distinct from zero.

Analyzing the results of the 9/11 discrimination analysis in isolation might present a puzzling picture. What could drive such a gap in behavior between inexperienced and inexperienced immigration judges? Insights from the learning curve exercise provide valuable context for the interpretation of these results. The knowledge that experienced judges are also the ones who are more mature in their judgement lends itself to the argument that a change in asylum-granting behavior towards Muslim asylees was guided by some form of discriminatory bias, such as the availability or representativeness heuristic, rather than objective deliberation. The change in grant behavior occurring along the lines of cognitive maturity also serves to discount the argument that decreasing grant rates for asylum applicants from Middle Eastern Arab countries was merely an appropriate calibration to an objectively higher threat from this treatment group.

Rather, this decline in grant rates is more reasonably attributed to the faulty and widespread incorporation of extraneous information into the calculus of whether a given asylum seeker deserved a grant. The salience of the September 11th attacks, and subsequent proliferation of media which directly (and indirectly) linked the Muslim community to notions of terrorism, promoted subconscious associations of danger with this group. Tragically, in affecting asylum seekers, this stigma affected even the most vulnerable and helpless members of this population.

My findings support the claim that the availability of harmful narratives after 9/11 influenced the decision process for inexperienced judges, exerting a material and devastating impact on many asylum seekers who would have been granted relief had their hearing occurred prior to the attacks. Furthermore, my paper warns more generally of disparate impact resulting from the susceptibility to cognitive biases. It would be a mistake to regard the judgement errors produced by cognitive biases as an unfortunate, but natural function of the decision process, which affect court decisions at random. Instead, these errors should be recognized as exceptionally harmful, as they serve to exacerbate existing patterns of racial and religious discrimination.

This research promotes several veins for future inquiry. First, there is much to be explored regarding the concept of the learning curve. In my research, I use the learning curve from a specific cognitive bias, the gambler's fallacy, to serve as a proxy for cognitive maturity upon which to base further analysis. As mentioned previously, however, nothing suggests that the gambler fallacy learning curve is wholly representative of the learning curve for other cognitive biases. Future research should explore the shape of learning curves with regard to a multitude of cognitive biases, as different shapes have different implications for when decision-makers reach a stagnation in their acquisition of cognitive benefits over time.

Furthermore, heterogeneity within learning curves should be explored. For example, do women "learn faster" than men? That is, do women exhibit steeper curves? Are there curvature differences among people from different socioeconomic backgrounds? Answers to these questions may offer illuminating insights with regard to the way cognition interacts with more permanent or external characteristics, and potentially inform the targeting of policies meant to combat the influence of cognitive biases.

Another area of consideration might be the way learning curves change across decision environments. For example, asylum adjudications can be characterized as relatively unstructured decision environments – judges get virtually no feedback, and base their decision off of very little substantial evidence. In comparison, one could argue that bail judges operate in an environment with relatively greater structure. Bail judges *can* get feedback, particularly on their decision to allow a defendant to await trial outside of jail, in the form of failures to appear, and recidivism. Their decisions are based not only on interactions at a hearing, but on criminal records, which can prove very informative for the prediction of future criminal activity. Do bail judges face a different learning curve than immigration judges? Perhaps their learning curve has a different "intercept", meaning that there are lower baseline levels of cognitive bias-driven behavior due to the higher level of structure, and therefore greater availability of signal, in their cases. Perhaps it has a different slope as well. Such points are important to explore in order to validate the appropriate application of a learning curve theory beyond the asylum hearing setting.

References

Abadie, Alberto. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects" Journal of Economic Literature (2019).

Abadie, Alberto, Alberto Abadie, and Javier Gardeazabal. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93, no. 1 (March 2003): 113–32.

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge. "When Should You Adjust Standard Errors for Clustering?" *NBER Working Paper No. 24003*, November 2017. https://www.nber.org/papers/w24003.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105, no. 490 (June 2010): 493–505. https://doi.org/10.1198/jasa.2009.ap08746.

Abrams, David S., Marianne Bertrand, and Sendhil Mullainathan. "Do Judges Vary in Their Treatment of Race Conference: The Law and Economics of Race." *Journal of Legal Studies* 41, no. 2 (2012): 347–84.

Arab American Institute. *Healing the Nation : The Arab American Experience after September 11*, 2002.

Arnold, David, Will Dobbie, and Crystal S. Yang. "Racial Bias in Bail Decisions." *The Quarterly Journal of Economics* 133, no. 4 (November 1, 2018): 1885–1932. https://doi.org/10.1093/qje/qjy012.

TRAC Immigration. "Asylum Decisions." Accessed April 10, 2020. https://trac.syr.edu/phptools/immigration/asylum/.

"Asylum Decisions and Denials Jump in 2018." TRAC Syracuse, November 29, 2018. https://trac.syr.edu/immigration/reports/539/.

Baicker, Katherine, and Theodore Svoronos. "Testing the Validity of the Single Interrupted Time Series Design." *SSRN Electronic Journal*, January 2019. https://www.nber.org/papers/w26080.

Basant, Rakesh, and Gitanjali Sen. "Impact of Affirmative Action in Higher Education for the Other Backward Classes in India." Working Paper. Indian Institute of Management Ahmedabad, July 18, 2016. http://vslir.iima.ac.in:8080/xmlui/handle/11718/20233.

Bergen, Peter, Albert Ford, Alyssa Sims, and David Sterman. "Who Are the Terrorists?" New America, 2018. https://www.newamerica.org/in-depth/terrorism-in-america/who-are-terrorists/.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics* 119, no. 1 (February 2004): 249–75.

Billmeier, Andreas, and Tommaso Nannicini. "Assessing Economic Liberalization Episodes: A Synthetic Control Approach." *Review of Economics and Statistics* 95, no. 3 (July 2013): 983–1001. https://doi.org/10.1162/REST_a_00324.

Bonanno, George A., and John T. Jost. "Conservative Shift Among High-Exposure Survivors of the September 11th Terrorist Attacks." *Basic and Applied Social Psychology* 28, no. 4 (n.d.).

Bun, Maurice J. G., and Teresa D. Harrison. "OLS and IV Estimation of Regression Models Including Endogenous Interaction Terms." *Econometric Reviews* 38, no. 7 (August 9, 2019): 814–27. https://doi.org/10.1080/07474938.2018.1427486.

Butcher, Kristin F., and Anne Morrison Piehl. "Cross-city Evidence on the Relationship between Immigration and Crime." *Journal of Policy Analysis and Management* 17, no. 3 (January 6, 1999).

Byers, Bryan D., and James A. Jones. "The Impact of the Terrorist Attacks of 9/11 on Anti-Islamic Hate Crime." *Journal of Ethnicity in Criminal Justice* 5, no. 1 (February 1, 2007): 43–56. https://doi.org/10.1300/J222v05n01_03.

Cainkar, Louise. "Targeting Muslims, at Ashcroft's Discretion." *Middle East Report Online*, March 14, 2003.

Camacho, Nuno, Bas Donkers, and Stefan Stremersch. "Predictably Non-Bayesian: Quantifying Salience Effects in Physician Learning About Drug Quality." *Marketing Science* 30, no. 2 (March 2011): 305–20. https://doi.org/10.1287/mksc.1100.0624.

Chalfin, Aaron. "What Is the Contribution of Mexican Immigration to U.S. Crime Rates? Evidence from Rainfall Shocks in Mexico." *American Law and Economics Review* 6, no. 1 (Spring 2014): 220–268.

Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue. "Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *The Quarterly Journal of Economics* 131, no. 3 (August 1, 2016): 1181–1242. https://doi.org/10.1093/qje/qjw017.

Chen, Daniel Li. "This Morning's Breakfast, Last Night's Game: Detecting Extraneous Influences on Judging." *IAST Working Papers*, 2016, 16–49.

Chen, Ming Hsu. “Explaining Disparities in Asylum Claims.” *Georgetown Public Policy Review* 12 (2006): 29.

Cooley, John K. “The Contagion Spreads:: The Assault on America.” In *Unholy Wars*, 3rd ed., 193–226. Afghanistan, America and International Terrorism. Pluto Press, 2002. https://doi.org/10.2307/j.ctt18fscwr.17.

Costalli, Stefano, Luigi Moretti, and Constantino Pischedda. “The Economic Costs of Civil War: Synthetic Counterfactual Evidence and the Effects of Ethnic Fractionalization.” *Journal of Peace Research* 54, no. 1 (2017): 80–98.

Croskerry, Pat. “The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them.” *Academic Medicine* 78, no. 8 (August 2003): 775–780.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. “Extraneous Factors in Judicial Decisions.” *Proceedings of the National Academy of Sciences* 108, no. 17 (April 26, 2011): 6889–92. https://doi.org/10.1073/pnas.1018033108.

“Despite Hiring, Immigration Court Backlog and Wait Times Climb.” TRAC Syracuse, May 15, 2017.

Donohue, John J., Abhay Aneja, and Kyle D. Weber. “Right-to-Carry Laws and Violent Crime: A Comprehensive Assessment Using Panel Data and a State-Level Synthetic Control Analysis.” *Journal of Empirical Legal Studies* 16, no. 2 (June 2019): 198–247.

Eagly, Ingrid, and Steven Shafer. “A National Study of Access to Counsel in Immigration Court.” *University of Pennsylvania Law Review* 164, no. 1 (December 2015): 1–91.

Englich, Birte, Thomas Mussweiler, and Fritz Strack. “Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts’ Judicial Decision Making.” *Personality and Social Psychology Bulletin* 32, no. 2 (February 1, 2006): 188–200. https://doi.org/10.1177/0146167205282152.

Eren, Ozkan, and Naci Mocan. “Emotional Judges and Unlucky Juveniles.” *American Economic Journal: Applied Economics* 10, no. 3 (July 2018): 171–205. https://doi.org/10.1257/app.20160390.

Federal Bureau of Investigation. “Crime in the United States 2001,” 2001.

Foner, Nancy. *Wounded City: The Social Impact of 9/11 on New York City*. Russell Sage Foundation, 2005.

Fragomen, Austin T. “The Illegal Immigration Reform and Immigrant Responsibility Act of 1996: An Overview.” *International Migration Review* 31, no. 2 (June 1, 1997): 438–60. https://doi.org/10.1177/019791839703100208.

"FY 2016 Statistics Yearbook." U.S. Department of Justice Executive Office for Immigration Review, March 2017.

Gallup Inc. "American Public Opinion and the Holocaust." Gallup.com, April 23, 2018. https://news.gallup.com/opinion/polling-matters/232949/american-public-opinion-holocaust.aspx.

Gold, Eric, and Gordon Hester. "The Gambler's Fallacy and the Coin's Memory." In *Rationality and Social Responsibility: Essays in Honor of Robyn Mason Dawes*, 21–46. Modern Pioneers in Psychological Science: An APS-Psychology Press Series. New York, NY, US: Psychology Press, 2008.

Goodman, J. David, and Ron Nixon. "Obama to Dismantle Visitor Registry Before Trump Can Revive It." *The New York Times*, December 22, 2016.

Grogger, Jeffrey. "Soda Taxes and the Prices of Sodas and Other Drinks: Evidence from Mexico." *American Journal of Agricultural Economics* 99, no. 2 (March 2017): 481–98. https://doi.org/10.1093/ajae/aax024.

Gut, Peter, and Stephen Jarrell. "Silver Lining on a Dark Cloud: The Impact of 9/11 on a Regional Tourist Destination." *Journal of Travel Research* 46, no. 2 (November 2007): 147–53.

Guthrie, Chris, Jeffrey J. Rachlinski, and Andrew J. Wistrich. "Blinking on the Bench: How Judges Decide Cases." *Cornell Law Review* 93, no. 1 (2008 2007): 1–44.

———. "Inside the Judicial Mind." *Cornell Law Review* 86, no. 4 (2001 2000): 777–830.

Hall, Katherine H. "Reviewing Intuitive Decision-Making and Uncertainty: The Implications for Medical Education." *Medical Education* 36, no. 3 (2002): 216–24. https://doi.org/10.1046/j.1365-2923.2002.01140.x.

Harris, Lindsay M. "The One-Year Bar to Asylum in the Age of the Immigration Court Backlog." *Wisconsin Law Review* 2016, no. 6 (2016): 1185–1250.

Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold Pollack. "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago." *The Quarterly Journal of Economics* 132, no. 1 (February 2017).

NPR.org. "How The Trump Administration's 'Zero Tolerance' Policy Changed The Immigration Debate." Accessed April 17, 2020. https://www.npr.org/2019/06/20/734496862/how-the-trump-administrations-zero-tolerance-policy-changed-the-immigration-deba.

Human Rights Watch. *"We Are Not the Enemy": Hate Crimes Against Arabs, Muslims, and Those Perceived to Be Arab Or Muslim After September 11*, n.d.

Ibish, H. “Report on Hate Crimes and Discrimination against Arab Americans: The Post-9/11 Backlash, 9/11, 2001-October 11, 2002. Washington, DC.” Washington, DC: American-Arab Anti-Discrimination Committee, 2003.

“ICEwatch: ICE Raids Tactics Map.” Immigrant Defense Project, July 2018.

Jones, Damon, and Ioana Marinescu. “The Labor Market Impacts of Universal and Permanent Cash Transfers: Evidence from the Alaska Permanent Fund.” Working Paper. Working Paper Series. National Bureau of Economic Research, February 2018. https://doi.org/10.3386/w24312.

Kahneman, Daniel. *Thinking, Fast and Slow*. Macmillan, 2011.

Keith, L.c., and J.s. Holmes. “A Rare Examination of Typically Unobservable Factors in US Asylum Decisions.” *Journal of Refugee Studies* 22, no. 2 (01 2009): 224–41. https://doi.org/10.1093/jrs/fep008.

Kennedy, Elizabeth G, Alison Parker, and Human Rights Watch (Organization). *Deported to Danger: United States Deportation Policies Expose Salvadorans to Death and Abuse*, 2020. https://www.hrw.org/sites/default/files/report_pdf/elsalvador0220_web_0.pdf.

Kenney, David Ngaruri, and Philip G. Schrag. *Asylum Denied: A Refugee’s Struggle for Safety in America*. Univ of California Press, 2009.

Keyes, Elizabeth. “Beyond Saints and Sinners: Discretion and the Need for New Narratives in the U.S. Immigration System.” *Georgetown Immigration Law Journal* 26, no. 2 (2012 2011): 207–56.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. “Human Decisions and Machine Predictions.” *The Quarterly Journal of Economics* 133, no. 1 (February 1, 2018): 237–93. https://doi.org/10.1093/qje/qjx032.

Lee, Sangkwon, Chi-Ok Oh, and Joseph T. O’Leary. “Estimating the Impact of the September 11 Terrorist Attacks on the US Air Transport Passenger Demand Using Intervention Analysis.” *Tourism Analysis* Vol. 9, no. No. 4 (2005): 355-361(7).

Mensah, Justice Tei, and Kweku Opoku-Agyemang. “Innocent Until Stereotyped Guilty? Terrorism and US Immigration Court Decisions,” December 2018.

Misra, Tanvi. “DOJ Changed Hiring to Promote Restrictive Immigration Judges.” Roll Call. Accessed April 13, 2020. https://www.rollcall.com/2019/10/29/doj-changed-hiring-to-promote-restrictive-immigration-judges/.

National Association of Immigration Judges. Improving Efficiency and Ensuring Justice in the Immigration Court System, § Senate Committee on the Judiciary (2011).

Nisbett, Richard E., David H. Krantz, Christopher Jepson, and Ziva Kunda. "The Use of Statistical Heuristics in Everyday Inductive Reasoning." *Psychological Review* 90, no. 4 (1983): 339–63. https://doi.org/10.1037/0033-295X.90.4.339.

Nizalova, Olena Y., and Irina Murtazashvili. "Exogenous Treatment and Endogenous Factors: Vanishing of Omitted Variable Bias on the Interaction Term." *Journal of Econometric Methods* 5, no. 1 (January 1, 2016). https://doi.org/10.1515/jem-2013-0012.

Norris, Samuel. "Judicial Errors: Evidence from Refugee Appeals." *SSRN Electronic Journal*, 2018. https://doi.org/10.2139/ssrn.3267611.

Penn State Law Immigrants' Rights Clinic and Rights Working Group. *The NSEERS Effect: A Decade of Racial Profiling, Fear, and Secrecy*. Center for Immigrants' Rights Clinic Publications, 2012. https://elibrary.law.psu.edu/irc_pubs/11/.

Philippe, Arnaud, and Aurélie Ouss. "'No Hatred or Malice, Fear or Affection': Media and Sentencing." *Journal of Political Economy* 126, no. 5 (October 2, 2018): 2134–78. https://doi.org/10.1086/699210.

Powell, Kimberly A. "Framing Islam: An Analysis of U.S. Media Coverage of Terrorism Since 9/11." *Communication Studies* 62, no. 1 (January 31, 2011): 90–112. https://doi.org/10.1080/10510974.2011.533599.

Pyun, Hyunwoong. "Exploring Causal Relationship between Major League Baseball Games and Crime: A Synthetic Control Analysis." *Empirical Economics* 57, no. 1 (July 1, 2019): 365–83. https://doi.org/10.1007/s00181-018-1440-9.

Rabby, Faisal, and William M. Rodgers. "Post 9-11 U.S. Muslim Labor Market Outcomes." *Atlantic Economic Journal* 39, no. 3 (July 26, 2011): 273. https://doi.org/10.1007/s11293-011-9281-3.

———. "The Impact of 9/11 and the London Bombings on the Employment and Earnings of U.K. Muslims." *IZA Discussion Paper No. 4763*, n.d.

Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Philip G. Schrag. "Refugee Roulette: Disparities in Asylum Adjudication Feature." *Stanford Law Review* 60, no. 2 (2008 2007): 295–412.

Rosenberg, Mica, Reade Levinson, and Ryan McNeill. "For U.S. Asylum Seekers, Some Judges Are a Better Bet than Others." Reuters. Accessed April 14, 2020. http://www.reuters.com/investigates/special-report/usa-immigration-asylum/.

Rydberg, Jason, Edmund F. McGarrell, Alexis Norris, and Giovanni Circo. "A Quasi-Experimental Synthetic Control Evaluation of a Place-Based Police-Directed Patrol Intervention on Violent Crime." *Journal of Experimental Criminology* 14 (February 8, 2018): 83–109.

Saunders, Jessica, Russell Lundberg, Anthony A. Braga, Greg Ridgeway, and Jeremy Miles. "A Synthetic Control Approach to Evaluating Place-Based Crime Interventions." *Journal of Quantitative Criminology* 31, no. 3 (September 1, 2015): 413–34. https://doi.org/10.1007/s10940-014-9226-5.

Shora, Kareem. "National Security Entry Exit Registration System (NSEERS) Future Issues: Speech." *Cardozo Public Law, Policy & Ethics Journal* 2, no. 1 (2004 2003): 73–80.

Sikorski, Christian von, Desiree Schmuck, and Jörg Matthes. "'Muslims Are Not Terrorists': Islamic State Coverage, Journalistic Differentiation Between Terrorism and Islam, Fear Reactions, and Attitudes Toward Muslims." *Mass Communication and Society* Vol 20, no. No 6 (August 4, 2017): Pages 825-848.

Skitka, L.J., C.W. Bauman, N.P. Aramovich, and G.S. Morgan. "Confrontational and Preventative Policy Responses to Terrorism: Anger Wants a Fight and Fear Wants 'Them' to Go Away." *Basic and Applied Social Psychology* 28, no. 4 (2006).

Thaler, Richard H., and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin, 2009.

Tversky, Amos, and Daniel Kahneman. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5, no. 2 (September 1, 1973): 207–32. https://doi.org/10.1016/0010-0285(73)90033-9.

———. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185, no. 4157 (1974): 1124–31.

U. S. Government Accountability Office. "Asylum: Variation Exists in Outcomes of Applications Across Immigration Courts and Judges," no. GAO-17-72 (November 14, 2016). https://www.gao.gov/products/GAO-17-72.

———. "U.S. Asylum System: Significant Variation Existed in Asylum Outcomes across Immigration Courts and Judges," no. GAO-08-940 (September 25, 2008). https://www.gao.gov/products/GAO-08-940.

"United States Deportation Policies Expose Salvadorans to Death and Abuse." Human Rights Watch, February 5, 2020. https://www.hrw.org/report/2020/02/05/deported-danger/united-states-deportation-policies-expose-salvadorans-death-and.

Vaala, Lindsey R. "Bias on the Bench: Raising the Bar for U.S. Immigration Judges to Ensure Equality for Asylum Seekers Note." *William and Mary Law Review* 49, no. 3 (2008 2007): 1011–42.

Yeh, Rick Fang-Chi. "Today's Immigration Legal System: Flaw and Possible Reforms." *Rutgers Race & The Law Review* 10, no. 2 (2009 2008): 441–68.

Appendix

Logit Regression *(Table 8)*	
	grant
Lag Grant*1 Year Experience	-0.196*** (0.047)
Lag Grant*2 Years Experience	-0.136*** (0.039)
Lag Grant*3 Years Experience	-0.096** (0.039)
Lag Grant*4 Years Experience	-0.070* (0.041)
Lag Grant*5 Years Experience	-0.123*** (0.042)
Lag Grant*6 Years Experience	-0.108*** (0.042)
Lag Grant*7 Years Experience	-0.114*** (0.042)
Lag Grant*8 Years Experience	-0.124*** (0.041)
Lag Grant*9 Years Experience	-0.106** (0.041)
Lag Grant*10 Years Experience	-0.104** (0.043)
Lag Grant*11 Years Experience	0.030 (0.045)
Lag Grant*12 Years Experience	0.003 (0.047)
Lag Grant*13 Years Experience	-0.091* (0.050)
Lag Grant*14 Years Experience	0.057* (0.033)
Lag Grant*>15 Years Experience	-
N	345,083
Log Likelihood	-156,021.300
Akaike Inf. Crit.	312,468.700
Notes:	***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Probit Regression *(Table 9)*	
	grant
Lag Grant*1 Year Experience	-0.101*** (0.027)
Lag Grant*2 Years Experience	-0.094*** (0.023)
Lag Grant*3 Years Experience	-0.069*** (0.022)
Lag Grant*4 Years Experience	-0.057** (0.023)
Lag Grant*5 Years Experience	-0.085*** (0.024)
Lag Grant*6 Years Experience	-0.072*** (0.024)
Lag Grant*7 Years Experience	-0.078*** (0.025)
Lag Grant*8 Years Experience	-0.121*** (0.024)
Lag Grant*9 Years Experience	-0.069*** (0.024)
Lag Grant*10 Years Experience	-0.064*** (0.025)
Lag Grant*11 Years Experience	0.018 (0.026)
Lag Grant*12 Years Experience	0.004 (0.027)
Lag Grant*13 Years Experience	-0.057 (0.039)
Lag Grant*14 Years Experience	0.059 (0.037)
Lag Grant*15 Years Experience	-
N	345,083
Log Likelihood	-155,766.000
Akaike Inf. Crit.	311,957.900
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

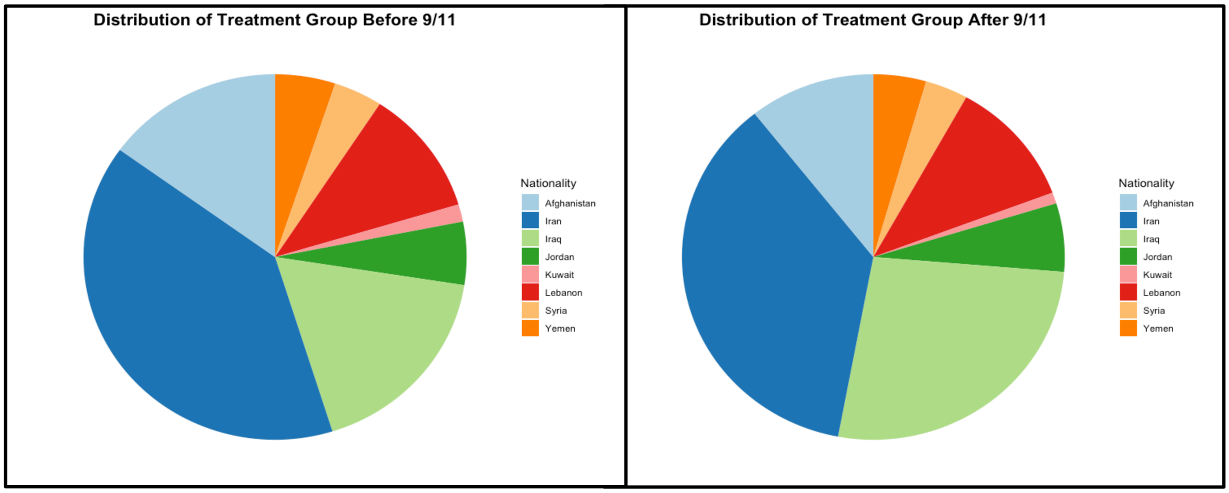
Alternate Dummy Baseline Levels *(Table 10)*

	Grant					
Baseline Dummy Level	>10 Years	>11 Years	>12 Years	>13 Years	>15 Years	>16 Years
Lag Grant:1st Year	-0.074*** (0.010)	-0.074*** (0.010)	-0.074*** (0.010)	-0.076*** (0.010)	-0.077*** (0.010)	-0.080*** (0.011)
Lag Grant:2nd Year	-0.056*** (0.009)	-0.056*** (0.009)	-0.056*** (0.009)	-0.058*** (0.009)	-0.059*** (0.009)	-0.062*** (0.009)
Lag Grant:3rd Year	-0.046*** (0.010)	-0.046*** (0.010)	-0.046*** (0.010)	-0.048*** (0.010)	-0.050*** (0.010)	-0.052*** (0.010)
Lag Grant:4th Year	-0.037*** (0.010)	-0.037*** (0.010)	-0.037*** (0.010)	-0.038*** (0.010)	-0.040*** (0.010)	-0.043*** (0.010)
Lag Grant:5th Year	-0.042*** (0.010)	-0.042*** (0.010)	-0.042*** (0.010)	-0.043*** (0.010)	-0.045*** (0.010)	-0.048*** (0.010)
Lag Grant:6th Year	-0.037*** (0.010)	-0.037*** (0.010)	-0.037*** (0.010)	-0.038*** (0.010)	-0.040*** (0.010)	-0.043*** (0.010)
Lag Grant:7th Year	-0.032*** (0.011)	-0.032*** (0.011)	-0.032*** (0.011)	-0.034*** (0.011)	-0.036*** (0.012)	-0.038*** (0.012)
Lag Grant:8th Year	-0.041*** (0.010)	-0.041*** (0.010)	-0.041*** (0.010)	-0.043*** (0.010)	-0.044*** (0.011)	-0.047*** (0.011)
Lag Grant:9th Year	-0.022** (0.010)	-0.022** (0.010)	-0.022** (0.010)	-0.024** (0.010)	-0.025** (0.011)	-0.028*** (0.011)
Lag Grant:10th Year	-0.024** (0.010)	-0.024** (0.010)	-0.024** (0.010)	-0.025** (0.011)	-0.027** (0.012)	-0.030*** (0.012)
Lag Grant:11th Year		-0.003 (0.010)	-0.003 (0.010)	-0.005 (0.011)	-0.006 (0.011)	-0.009 (0.012)
Lag Grant:12th Year			-0.001 (0.010)	-0.003 (0.011)	-0.005 (0.012)	-0.007 (0.012)
Lag Grant:13th Year				-0.018 (0.012)	-0.020 (0.013)	-0.022* (0.013)
Lag Grant:14th Year					0.011 (0.012)	0.008 (0.013)
Lag Grant:15th Year					-0.021* (0.012)	-0.016 (0.012)
Lag Grant:16th Year						-0.019 (0.013)
N	345,083	345,083	345,083	345,083	345,083	345,083
R^2	0.374	0.374	0.374	0.375	0.375	0.375
Adjusted R^2	0.374	0.374	0.374	0.374	0.374	0.374
Residual Std. Error	0.383 (df = 348801)	0.383 (df = 348799)	0.383 (df = 348797)	0.383 (df = 348795)	0.383 (df = 348791)	0.383 (df = 348789)
F Statistic	1,023.536*** (df = 204; 348801)	1,013.698*** (df = 206; 348799)	1,003.969*** (df = 208; 348797)	994.555*** (df = 210; 348795)	976.187*** (df = 214; 348791)	967.179*** (df = 216; 348789)
Notes:						***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Nationality Distributions *(Table 11)*

Country	Before 9/11	After 9/11
Afghanistan	0.148	0.105
Bahrain	0.005	0.007
Iran	0.391	0.359
Iraq	0.174	0.262
Jordan	0.055	0.060
Kuwait	0.015	0.010
Lebanon	0.109	0.110
Oman	0.001	-
Qatar	0.002	-
Saudi Arabia	0.006	0.008
Syria	0.040	0.036
United Arab Emirates	-	0.001
Yemen	0.050	0.044

Figure 13: Comparison of Treatment Group's Nationality Composition



Distribution of Treatment Group Before 9/11

Distribution of Treatment Group After 9/11

Political Party vs Experience Group *(Table 12)*

	Inexperienced	Experienced	Total
Republican-appointed	56	55	111
Democrat-appointed	142	10	152
Total	198	65	263

Chi-Squared Statistic: 63.6629
P-value: < .00001

Figure 14: Inexperienced Gap Plot-Controls with MSPE limit of 10x

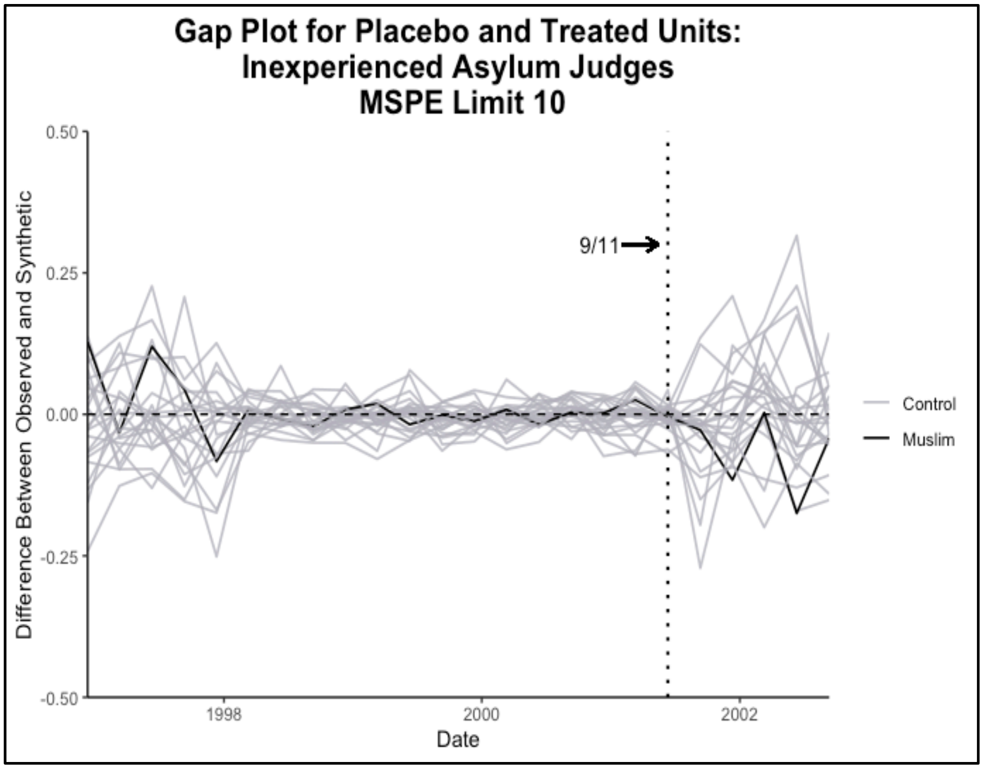


Figure 15: Inexperienced MSPE Ratio Plot- Controls with MSPE limit of 10x

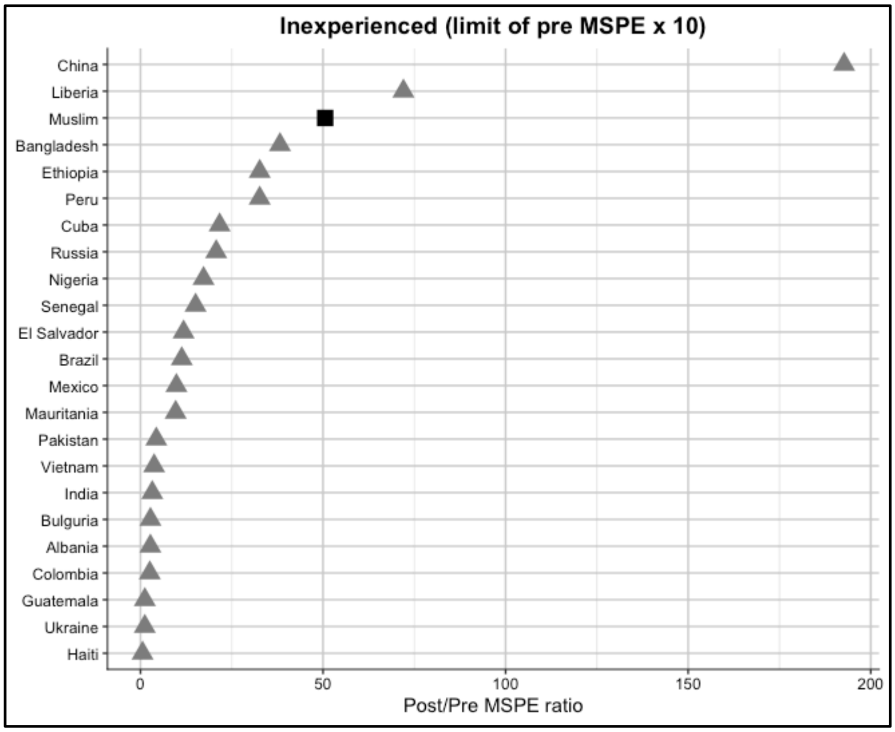


Figure 16: Experienced Gap Plot – Controls with MSPE limit of 10x

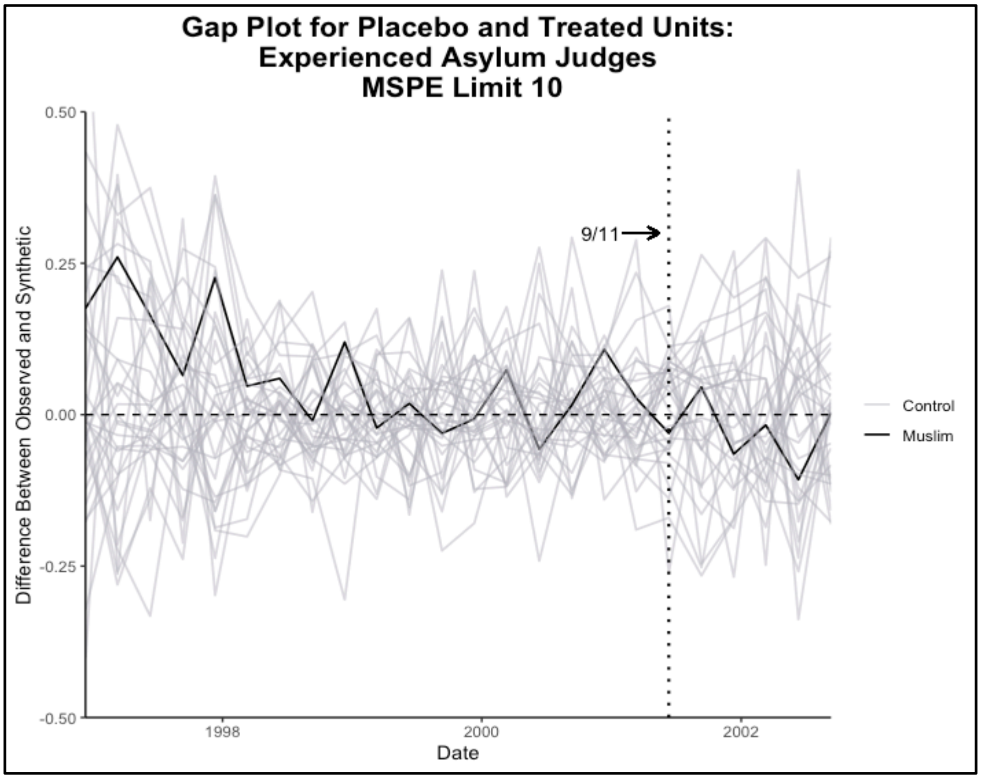


Figure 17: Experienced MSPE Ratio Plot – Controls with MSPE limit of 10x

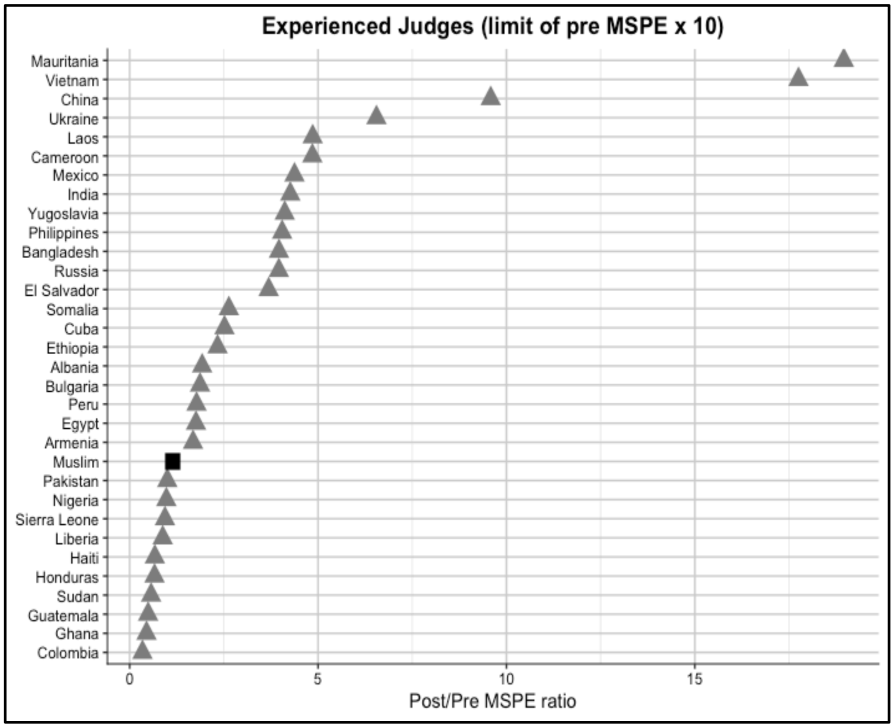


Figure 18: Inexperienced Gap Plot – Controls with MSPE limit of 20x

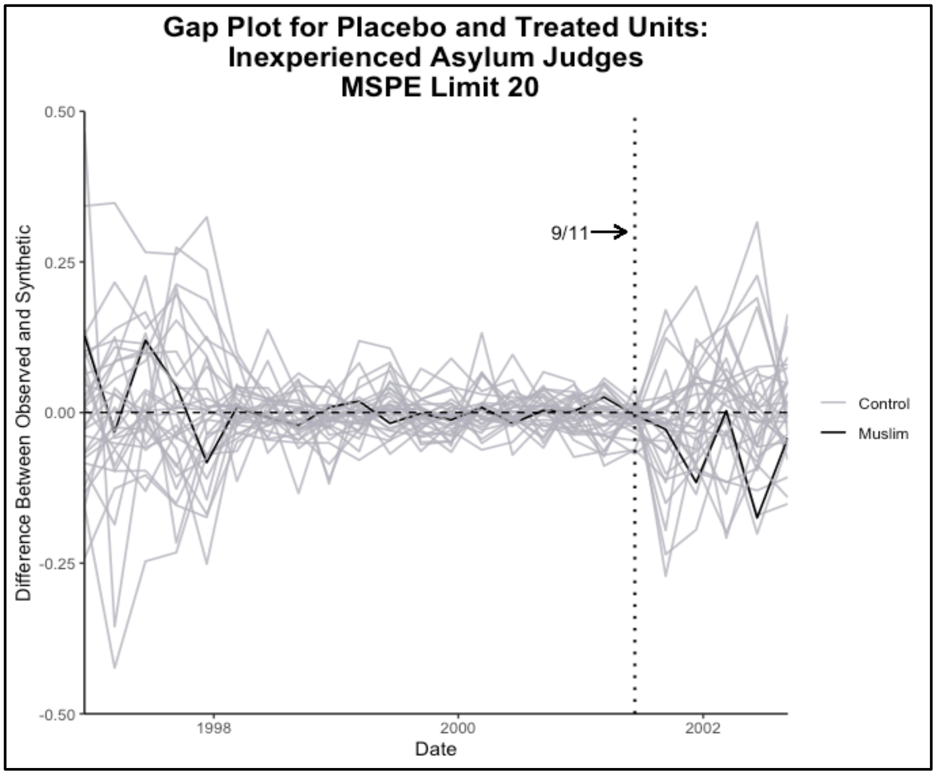


Figure 19: Inexperienced MSPE Ratio Plot – Controls with MSPE limit of 20x

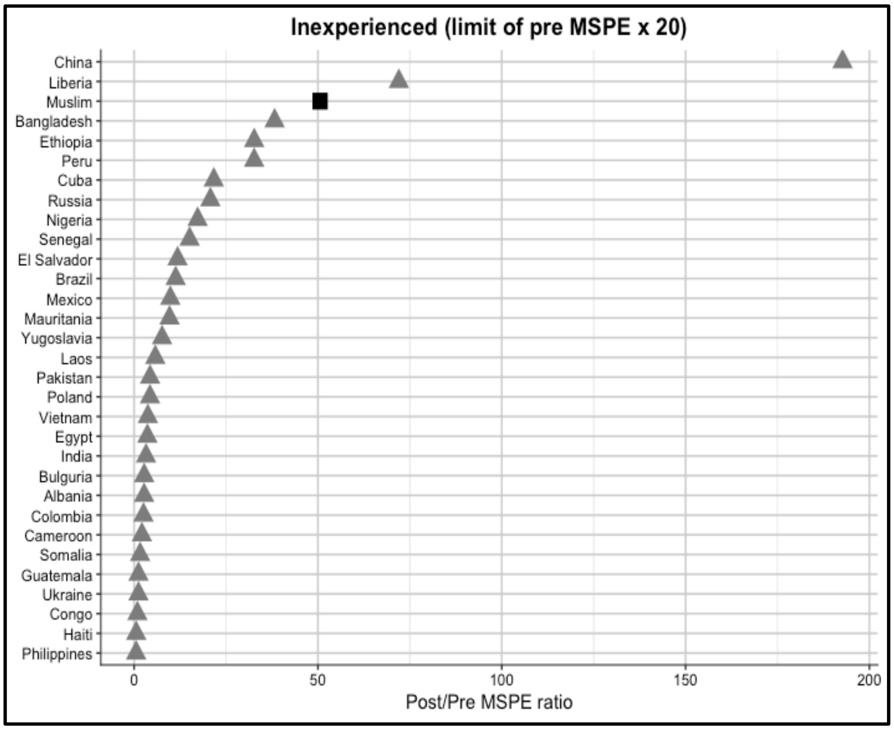


Figure 20: Experienced Gap Plot – Controls with MSPE limit of 2x

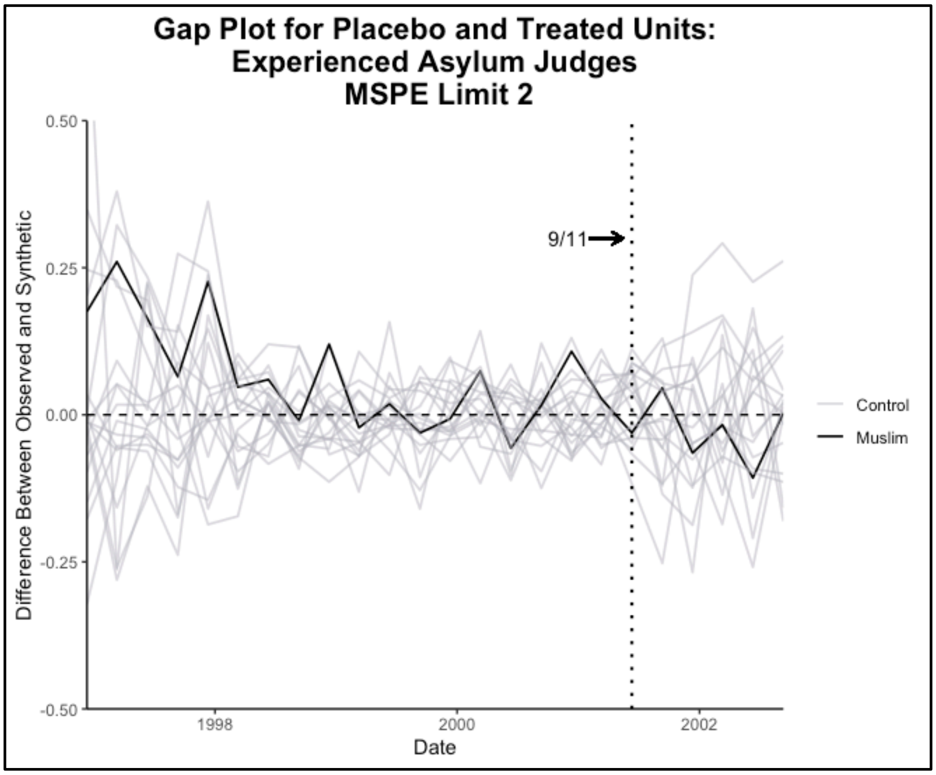
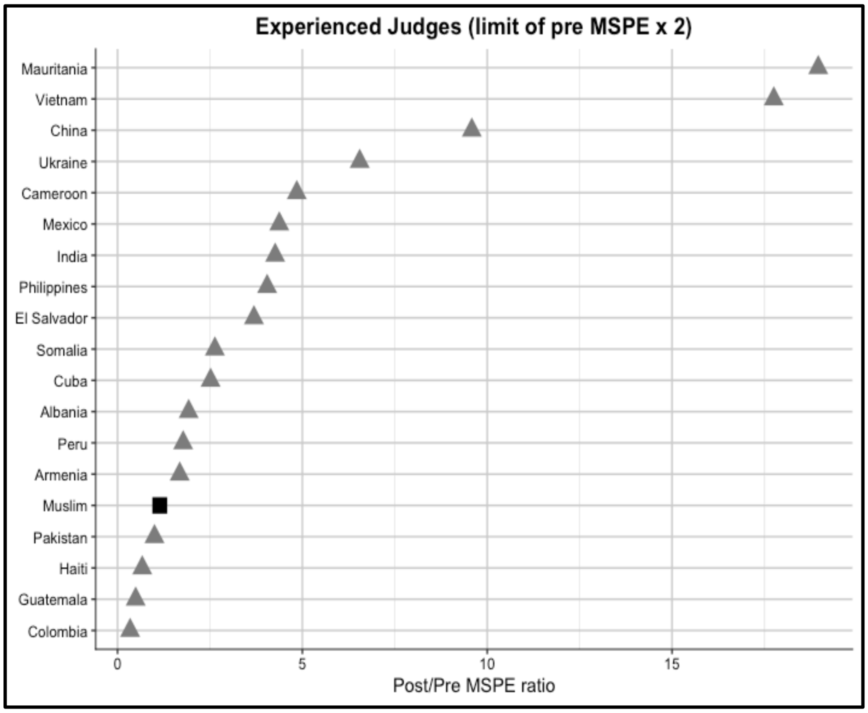


Figure 21: Experienced MSPE Ratio Plot – Controls with MSPE limit of 2x


Experienced Judges (limit of pre MSPE x 2)
Mauritania
Vietnam
China
Ukraine
Cameroon
Mexico
India
Philippines
El Salvador
Somalia
Cuba
Albania
Peru
Armenia
Muslim
Pakistan
Haiti
Guatemala
Colombia
Post/Pre MSPE ratio