Judicial Learning from Feedback: Evidence from U.S. Asylum Courts *

Sai Zhang
USC

Daniel L. Chen
Toulouse

March 23, 2026

[Click for the latest version]

Abstract

We study how immigration judges learn from appeal feedback in a setting with substantial discretion and noisy feedback. Using comprehensive EOIR and BIA administrative data, we construct high-dimensional prediction models to proxy judges' expected reversal probabilities and identify surprise reversals where the appeal outcome sharply deviates from judicial expectations. Event-study estimates show that surprise reversals increase grant rates by 1.57 percentage points in the following week, with no evidence of reduced procedural diligence. An attention-based learning model rationalizes these dynamics, predicting that only attentive judges update strongly in response to unexpected feedback. Reduced-form and Marginal Treatment Effect (MTE) analyses confirm this heterogeneity: attentive judges converge to appellate standards, while inattentive judges remain threshold-misaligned. Learning operates through both appeal filing and appeal success, and is stronger in high-salience defensive cases. Our findings highlight attention as a key channel through which appeal courts effectively discipline lower-court decision-making.

*This paper has greatly benefited from suggestions of Jeff Weaver, Vittorio Bassi and Alex Imas. We thank Viknesh Nagarathinam and Lilian Yang for outstanding research assistance. Zhang: University of Southern California, saizhang@usc.edu. Chen: Toulouse School of Economics, daniel.chen@iast.fr.

Before an Arrest, Officers Tossed a (Virtual) Coin.

New York Times, July 14, 2018

1 Introduction

High stake judicial decisions are routinely made in environments with uncertainty, heterogeneous expertise, and noisy or delayed feedback. This is particularly apparent than in the U.S. asylum system, where immigration judges determine life altering outcomes that ultimately shape humanitarian protection, migration flows, and the credibility of legal institutions. Asylum judges operate with broad discretion: conditioning on case characteristics, grant rates vary by more than 40 percentage points across judges on identical dockets, a level of dispersion rare among expert decision makers. Despite the high policy stakes, the appeal process does not provide strong feedback since most decisions judges make cannot easily be categorized as correct or incorrect absent an extreme ruling (Rachlinski and Wistrich, 2017). In addition, the Board of Immigration Appeals reviews lower court decisions often months after the original ruling, and its opinions vary dramatically in informativeness. As a result, the institutional environment resembles a low feedback and high discretion system, precisely the type where cognitive biases, heuristic decision rules, and non-Bayesian updating tend to present (Rachlinski and Wistrich, 2017).

This setting raises a question: How can judges improve or calibrate their decision standards when feedback is delayed and ambiguous? Classic principal agent theories predict compliance with appeal preferences, yet substantial and persistent disparities remain. Existing empirical studies document that lower courts sometimes respond to reversals, but the mechanisms remain unclear, and the literature lacks causal evidence on whether judges actually learn from appeal results or simply react strategically to sanctions. Immigration courts provide an ideal environment to address this question because the institutional structure generates large natural variation in appeal signals, and the consequences of error are salient enough that true learning, rather than strategic signaling, should matter.

In this paper, we provide novel causal evidence on how immigration judges update their decision standards in response to reversals, defined as cases denied in the lower court but granted on appeal. Leveraging comprehensive EOIR pro-

ceeding level data, BIA appellate rulings, lower court judge biographies, and high dimensional case features, we construct separate random forest prediction models for both courts. These models achieve 76.7% out-of-sample accuracy for lower court decisions and 74.8% accuracy for appeal decisions. The appeal prediction model uses only information available to lower-court judges at the time of ruling, it serves as a proxy for judges' internal expectations of appeal success. We use it to isolate surprise reversals, the cases that the model predicts as affirmed denials but are instead overturned. These reversals represent appeal outcomes that deliver highly informative, belief updating signals to judges, allowing us to approximate the information shocks judges plausibly receive.

We design a weekly event-study framework to examine how judges react to these surprise reversals. It reveals two main results: First, judges increase grant rates by 1.57 percentage points in the first week following a surprise reversal, and this shift persists for several weeks thereafter. Second, judges show no evidence of reduced diligence: rulings do not become more predictable using predetermined features before case opening, hearing durations do not shorten, written decisions do not shrink, and other observable effort margins are unchanged. These patterns rule out mechanical substitution away from effort and instead point to a shift in attention to evaluation signals.

To interpret these results, we develop a simple attention based learning model in which judges selectively attend to feedback and update their granting thresholds after reversals revealing new information about the appeal court standard. The model predicts a sharp distinction between attentive and inattentive judges: attentive judges place greater weight on surprise reversals and converge to the appeal standard, while inattentive judges update weakly and remain "sticky" in their granting thresholds. This mechanism aligns with theories of cognitive frictions, limited attention, and endogenous information acquisition (Spitzer and Talley, 2013; Buser, Gerhards and Van Der Weele, 2018).

We then bring the model predictions to the data using both the event study design and the leave-out judge leniency instrument design (Goldsmith-Pinkham, Hull and Kolesár, 2025). We classify judges into attentive and inattentive types based on their behavioral response to surprise reversals and test the model's implications through reduced-form and marginal treatment effect (MTE) analyses (Heckman and Vytlacil, 2005; Heckman, Urzua and Vytlacil, 2006). We show that

attentive judges converge to appeal court standards and show near zero marginal reversal elasticity: marginal shifts in their leniency no longer meaningfully affect reversal probabilities. Inattentive judges, by contrast, display persistent threshold misalignment, with a large density of borderline cases whose outcomes remain highly sensitive to small changes in their personal decision thresholds.

We further decompose reversal rates into the extensive margin (appeal filing among denied cases) and the intensive margin (appeal grant conditional on appeal). Both margins contribute to the observed learning patterns, consistent with appellate feedback affecting both litigant behavior and appeal evaluation. Finally, we provide additional validation by comparing defensive cases, which generate stronger and more salient signals, to affirmative cases. Defensive cases induce convergence even among inattentive judges, whereas affirmative cases trigger meaningful updating only among attentive judges, reinforcing the centrality of attention in judicial learning.

Our results demonstrate that appeal feedback influences judicial behavior not through deterrence or effort substitution, but through selective attention to unexpected evaluation signals. Identifying this mechanism fills a critical gap in our understanding of how lower courts adapt to higher court oversight in low feedback environments. By distinguishing judges along an empirically validated attention margin, our framework provides a unifying explanation for the coexistence of persistent cross-judge disparities and systematic post-reversal adjustments.

Contribution to Literature Our study speaks to several literatures in law, economics, and political science.

First, we contribute to the literature on judicial responses to appeal court oversight. Existing work connects strategic compliance, precedent adherence, and error correction to the assumption that judges are averse to reversals and internalize the risk of reversals into their ruling (Cameron, Segal and Songer, 2000; Klein and Hume, 2003; Gennaioli and Shleifer, 2008; Westerland et al., 2010; Boyd, 2015; Chen, 2023). However, earlier studies lack causal evidence on the relationship between lower- and higher-court decisions, and few studies systematically examine the feedback embedded in appeal decisions. The closest study to ours is Bhuller and Sigstad (2025), who provide compelling causal evidence that criminal judges in Norway respond to reversals in a way consistent with Bayesian updating un-

der limited or selective memory. We differ in three major aspects: (i) we use random forest prediction models to identify surprise reversals, which convey effective belief-shifting information; (ii) we show that behavioral adjustments occur without reductions in procedural effort, pointing to attention-based rather than cost-shifting mechanisms; and (iii) leveraging a much larger sample, we develop a theoretical framework and provide empirical evidence on heterogeneous learning across the distribution of judge leniency.

Second, our paper relates to the large body of work on behavioral judicial decision-making. Recent research documents that judicial decisions are sensitive to irrelevant contextual factors (Eren and Mocan, 2018; Philippe and Ouss, 2018; Chen and Philippe, 2023; Heyes and Saberian, 2019), exhibit path dependence tendency (Chen, Moskowitz and Shue, 2016; Bindler and Hjalmarsson, 2019), and reflect group specific biases (Shayo and Zussman, 2011; Abrams, Bertrand and Mullainathan, 2012; Arnold, Dobbie and Yang, 2018; Kleinberg et al., 2018). We show that procedure type, defensive versus affirmative cases, is an additional dimension shaping judicial behavior, operating specifically through differences in attention and responsiveness to feedback. This suggests that the judicial hierarchy, although vague and delayed in its feedback, can still correct systematic biases when judges actively attend to the information it provides.

Third, we contribute to the emerging literature on endogenous information acquisition and selective attention in expert decision-making. A central insight in this literature is that decision-makers do not absorb all available information. Instead, they choose what to attend to based on cognitive costs, perceived benefits, and the informativeness of signals (Schwartzstein, 2014; Enke and Graeber, 2023; Bastianello and Imas, 2025). Field evidence shows that individuals update only when feedback is salient, surprising, or costly to ignore (Hanna, Mullainathan and Schwartzstein, 2014; Buser, Gerhards and Van Der Weele, 2018), and recent work highlights how professionals balance information processing costs against the risks of relying on outdated heuristics (Capozza et al., 2021). We provide rare causal evidence of attention and learning in a high-stakes institutional environment with noisy feedback, offering a framework for understanding when and why professionals learn from noisy evaluative signals.

Structure of the Paper The remainder of the paper is organized as follows: Section 2 describes the institutional setting of the U.S. asylum system and the data used in our analysis. Section 3 outlines the construction and implementation of the random forest prediction models for lower and appeal court decisions. Section 4 presents the event-study design and estimates the causal impact of surprise reversals on judicial behavior. Section 5 develops the attention-based learning framework and provides empirical tests of its key predictions. Section 6 concludes.

2 Background and Data

In this section, we introduce the institutional background of the U.S. immigration court system and describe the data for the analysis.

2.1 Background: US Immigration Court

The Executive Office for Immigration Review (EOIR) Under the United States Refugee Act of 1980, the United States may grant asylum to individuals who face persecution in their home countries on the basis of "race, religion, nationality, membership in a particular social group, or political opinion."[1] Many asylum seekers are already residing within the United States when they apply for protection. If an individual's application is denied by the U.S. Citizenship and Immigration Services (USCIS) Asylum Office, or if removal proceedings are initiated, they may renew their asylum claim before an Immigration Judge within the Executive Office for Immigration Review (EOIR).[2] A single judge presides over evidentiary hearings, determines removability, and adjudicates applications for relief such as asylum, withholding of removal, or protection under the Convention Against Torture. Decisions are based on testimony, supporting evidence, and the statutory framework outlined in the Immigration and Nationality Act (INA).

Within each local Immigration Court[3], cases are administratively assigned to individual judges. EOIR guidance provides that upon filing, "the case is randomly

[1] GovTrack.us. (2025). *S. 643 — 96th Congress: Refugee Act of 1979.* Retrieved from https://www.govtrack.us/congress/bills/96/s643.

[2] U.S. Department of Justice. *Immigration Court Practice Manual.* Retrieved from https://www.justice.gov/eoir/reference-materials/ic.

[3] Also referred to as the lower court hereafter.

assigned to a specific judge within that court," with subsequent motions generally decided by the assigned judges absent unavailability.[4] In practice, this institutional feature underpins empirical strategies that treat judge leniency as a plausibly quasi-random shifter of outcomes.[5] Noticably, each judge manages a high volume of cases while drafting reasoned written opinions. Internal performance evaluations emphasize both timeliness and quality of adjudication, generating tensions between throughput and legal precision. These operational constraints are central to understanding the behavioral environment in which judges make discretionary judgments.

The Board of Immigration Appeals (BIA) If asylum is denied, the applicant may appeal to the Board of Immigration Appeals (BIA), which exercises nationwide appellate jurisdiction.[6] The appellant must file a notice of appeal directly with the BIA within the prescribed deadline; appeals may not be filed with the Department of Homeland Security (DHS) or the Immigration Court. Appeals from the EOIR Immigration Court to the BIA are typically reviewed and decided by a single BIA member, though cases raising novel or complex legal issues may be assigned to a three-member panel.[7] The BIA reviews factual findings under a clear-error standard and reviews questions of law, discretion, or judgment *de novo*.[8] Given that initial asylum outcomes hinge on the judgment of a single decision-maker, the appeal process provides a critical safeguard, offering applicants a structured opportunity to seek a reversal of an asylum denial. BIA decisions themselves can be reviewed by the regional U.S. courts of appeals via petitions for review, which serve as an additional layer of judicial scrutiny.

[4]Office of the Chief Immigration Judge, OPPM 01-02: *Change of Venue* (Feb. 15, 2002), p. 1. Retrieved from https://www.justice.gov/sites/default/files/eoir/legacy/2002/02/15/OPPM01-02.pdf.

[5]TRAC, *Asylum Success Still Varies Widely Among Immigration Judges* (Nov. 25, 2024), discussing random assignment and persistent judge-level disparities. Retrieved from https://tracreports.org/reports/752/.

[6]EOIR, *Board of Immigration Appeals*. Retrieved from https://www.justice.gov/eoir/board-of-immigration-appeals.

[7]BIA Practice Manual (Composition of the Board). Retrieved from https://www.justice.gov/eoir/reference-materials/bia.

[8]8 C.F.R. § 1003.1(d)(3); American Immigration Council, *Standards of Review Applied by the BIA* (Apr. 22, 2020). Retrieved from https://www.americanimmigrationcouncil.org/wp-content/uploads/2025/01/standards_of_review_applied_by_the_board_of_immigration_appeals.pdf.

The incentive structure of EOIR judges The institutional design of EOIR and the two-tier appellate system creates well-defined incentives. Judges face two opposing risks. If they are too strict, denying relief in potentially meritorious cases, applicants are likely to appeal successfully. Reversals by the BIA not only increase case backlogs but may be perceived as signals of poor judgment. Conversely, if they are too lenient, granting asylum in marginal cases, they risk criticism from DHS, internal administrative review, or political oversight tied to aggregate approval rates—pressures that can be amplified during periods of heightened immigration enforcement. This dual threat leads judges to calibrate decisions carefully to produce decisions that are legally defensible, while disposing of cases efficiently to meet performance benchmarks.

For more details of the asylum adjudication process, we refer the interested reader to Ramji-Nogales, Schoenholtz and Schrag (2007).

2.2 Data

Our analysis employs multiple data sources linked through unique identifiers for each proceeding, case and judge.

EOIR data Following previous work (Chen, Moskowitz and Shue, 2016; Dunn et al., 2017), we use administrative records on U.S. refugee asylum cases considered in immigration courts from 1986 to 2019. The data provide comprehensive coverage of all asylum proceedings during this period. After pre-processing and data cleaning, the final sample comprises 832,946 unique court proceedings. Among these, 297,965 proceedings—representing 35.8% of all—resulted in a grant of asylum. For each proceeding, the dataset reports detailed information on the procedural timeline (application filing, case opening, and decision dates), adjudication outcomes, and court location, among other administrative attributes and case information. In Table A.1 and Table A.2 we present the ten most common countries of origin for asylum applicants and the ten U.S. states with the highest number of hearings, respectively.

We complement the proceeding-level data with a separate dataset of hearing-level records, which contains 24,725,971 individual hearing observations. Among these, 4,665,402 hearings correspond to cases in our main analytical sample. The

hearing-level data include the precise clock time of hearings and the recorded reasons for adjournments.

BIA data The raw BIA appeal dataset contains 1,108,046 appeal records in total. We apply several filters to construct the final analytical sample. First, we drop cases without a proceeding or case identifier, retaining 89.5% of the observations. Second, we restrict the sample to appeal types relevant to our analysis—*Appeal of IJ MTR*, *Case Appeal*, *Circuit Court Remand*, *Interlocutory Appeal*, and *MTR BIA*—leaving 478,006 observations. Third, for proceedings with multiple appeals, we retain only the most recent appeal, yielding 357,766 records. We then drop 52 appeals without valid filing year or case type information.

Following the BIA decision classifications summarized in Table A.3, we group the 27 distinct appeal outcomes into three categories: positive (granted), negative (dismissed), and neutral. Only appeals with positive or negative outcomes are retained for the predictive analysis, leaving 286,440 observations. In the final sample, out of 534,981 denied lower-court proceedings, 355,127 resulted in an appeal filing. Among these, 93,332 appeals were granted and 191,152 were denied.[9] Figure A.1 displays the distribution of lower-court and appellate decisions over the sample period.

Immigration judge biography data We hand-collected biographical information for 698 EOIR judges, including their year of appointment, educational background, and prior professional experience. Table A.4 summarizes selected characteristics of these judges.

3 Random-Forest Prediction of Asylum Outcomes

In this section, we outline the machine learning framework used to generate predictions of lower court and appeal court decisions in asylum cases. The objective is to develop an efficient model that predicts the binary outcome of asylum grant versus denial at both stages of adjudication.

[9] Due to the length of the appeals process, 54,330 cases filed in 2019 and 27,065 cases filed in 2018 remain undecided at the time of data extraction.

We employ random-forest based models as our primary predictive approach for two main reasons. First, most explanatory variables are categorical. When fully dummified, these features create a high-dimensional and sparse design matrix, often with substantial multicollinearity. Random forests handle such settings effectively without requiring variable selection or strong parametric assumptions (Breiman, 2001). Second, for binary classification problems, random forest models generally outperform alternative algorithms in terms of both predictive accuracy and robustness (Fernández-Delgado et al., 2014; Grinsztajn, Oyallon and Varoquaux, 2022).

For prediction tasks in both the lower court and the appeal court, we train and tune random-forest based models sequentially, using only data from cases filed in previous years to predict outcomes in subsequent years. This rolling design ensures that model estimation relies exclusively on information available at the time of decision making. Details of model specification and hyperparameter tuning are discussed in Section B.1.

3.1 Lower Court Prediction

For the lower-court decision-making process, our primary objective is to test how *early* judges' decisions can be predicted. By "early", we refer to predictions based solely on information predetermined before a case is opened and before any hearings are held. The goal is to evaluate how accurately such ex ante characteristics can explain actual adjudication outcomes.

To quantify the extent of early predictability, we estimate five models that progressively incorporate richer sets of predetermined variables. The analysis is based on 831,656 asylum proceedings, with the test sample defined as cases filed in each year from 1986 to 2019 and the corresponding training sample consisting of all cases from previous years. This rolling design extends the previous framework of Dunn et al. (2017), allowing the model to be re-estimated annually using only information available at the time of prediction. The five model specifications are as follows: (1) judge identifier only; (2) judge identifier and applicant nationality; (3) judge identifier and case opening date; (4) judge identifier, applicant nationality, and opening year-month; and (5) judge identifier, applicant nationality, opening year-month, and filing year-month.

As shown in Table A.5, the best-performing early-prediction model achieves an accuracy of 76.7%. This result indicates that, using only the predetermined information available at the time a case is opened, the model can anticipate judicial outcomes with remarkable accuracy. Such a high level of predictability underscores the extent to which pre-hearing characteristics already encapsulate much of the information that ultimately determines the final decision. We interpret this early-predictability measure as a proxy for how much judges' decisions rely on predetermined case attributes. It reflects the limited additional informational content extracted during the hearing process.

3.2 Appeal Court Prediction

For the appeal court prediction, our objective is to construct the most accurate predictive model possible. The model incorporates detailed case characteristics, information about the corresponding lower court judge, and time horizon features. Because all input variables are observable to lower court judges, the model effectively approximates the information set judges could use to form expectations about the likelihood of reversal, i.e., the probability that an appeal will be granted. Accordingly, the fine-tuned random-forest model can be interpreted as a best-case simulation of judges' beliefs regarding their own reversal risk.

For the appeal court cases, the model is estimated on 286,440 appeals with binary outcomes, either denial or grant. For each prediction year from 1986 to 2019, the model is trained on all appeals filed in previous years and tested on cases filed in the corresponding year.

The best-performing model achieves an overall accuracy of 74.8%. It performs particularly well in identifying appeal denials, with a true negative rate (TNR) of 90.0%, but is less effective in predicting appeal grants, with a true positive rate (TPR) of 43.6%. [10] Using the trained model to predict appeal outcomes for applicants who were denied in the lower court but did not file an appeal, we find that only 7.1% of such non-appealers would have been likely to succeed on appeal. Moreover, among those whom the model predicts would be granted an appeal, 82.8% in fact appealed. Figure A.2 plots the monthly time series of predicted out-

[10] The true positive rate (TPR) is the share of appeals correctly predicted as granted among all granted appeals, while the true negative rate (TNR) is the share of appeals correctly predicted as denied among all denied appeals.

comes, illustrating the stability of the model's predictive performance throughout the sample period. These results further confirm the model's predictive power and suggest that most non-appealers correctly assessed their low probability of appeal succeed. In Section B.2, we provide additional details on the appeal court prediction exercise.

4 Lower Court Judges' Response to Appeal Outcomes

From the perspective of a lower-court judge, a case that is denied at the lower level but later granted appeal constitutes an effective reversal. Therefore, a natural question is whether, and to what extent, lower-court judges judges adjust their subsequent decision making in response to such reversals. In this section, we address this question by employing a weekly event-study framework to analyze adjustments in judicial behavior.

4.1 Event Study Design

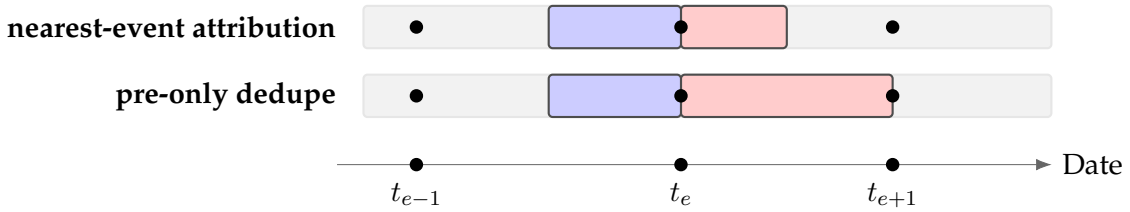
Formally, consider each appeal decision (*event*) e, with decision date t_e. We construct a symmetric event window of 13 weeks centered on the event, indexed by $k \in [-6, \ldots, 0, \ldots, 6]$, where each k corresponds to a seven-day period relative to t_e. Specifically, we define the first seven-day period following the appeal decision, $[t_e, t_e + 6]$, as the onset period of the reversal $k = 0$. For every lower-court case i decided by judge j on date t, we assign that case to a time window k relative to an event e experienced by the same judge. The resulting panel of case-event-period observations allows us to trace how judges' grant probabilities evolve before and after reversals. We define the event-study periods in seven-day intervals because, as shown in Figure A.3, the arrival probability of appeal decisions relative to lower-court rulings exhibits significant weekly periodicity. Aggregating the data to weekly periods yields a stable treatment probability across cases and over time, mitigating mechanical day-of-week fluctuations that could violate parallel trends.

Sample trimming A key challenge to our event-study design is that judges may experience multiple reversals within a relatively short period, such that the same

lower-court case could fall within multiple event windows. In particular, a lower-court case may be classified as pre-event for one reversal and post-event for another. In our preferred specification, we address this issue using the **nearest-event attribution** rule. As illustrated in Figure 1, each lower-court case is assigned to the appeal decision closest in time. This midpoint rule partitions the timeline by the halfway points between consecutive reversals, ensuring that no lower-court decision is simultaneously classified as both pre-event and post-event.

For robustness, we implement an alternative trimming method, **pre-only deduplication**. This approach maximizes sample size by retaining all post-event observations while deduplicating only on the pre-event side. Specifically, for cases falling into both pre-event windows and post-event windows, we keep only those closer to the current event than to the previous one as pre-event observations. No trimming is applied to the post-event side.

Figure 1: Event Study Sample Trimming Methods



Notes: The figure illustrates a stylized example of the two overlap-trimming rules, nearest-event trimming and pre-only dedupe trimming for the event study design described in Section 4. Dots represent the date of the current event e, the last event $e-1$ and the next event $e+1$, blue rectangle represents the pre-event sample kept in both methods, red rectangle represents the post-event sample kept in both methods.

For both methods, we retain lower-court cases that fall exclusively in either pre-event or post-event windows (pure pre- and pure post-observations). The nearest-event attribution method effectively eliminates overlapping contamination but reduces the estimation sample significantly, retaining 1,760,863 case-event-period observations (16% of the full sample). Meanwhile, the pre-only deduplication method retains 7,293,200 observations (66.3% of the full sample) but yields less precise estimates of post-event trends relative to the main trimming method.

Sample Balancing Figure A.4 plots the distribution of reversals kept in the sample after trimming with the nearest-event attribution method. The trimmed sample is concentrated within event windows $k \in [-2, 2]$. To improve balance across

periods, we aggregate observations from three or more weeks before and after an event into two broader bins in our main specification.

Another concern is that a lower-court case can appear in the pre-event windows of multiple events (or on the post-event side), giving such observations disproportionate influence in the estimation. In addition, some events contain more lower-court observations than others and disproportionately affect the estimates. If outcomes or dynamics differ systematically between large and small events, the unweighted estimates may reflect that heterogeneity in a biased way. For this, we also estimate an event-weighted specification. Specifically, for case i associated with event e at event time k, we calculate the weight $w_{ie}(k) = \frac{1/N_{e,k}}{\sum_{e'} 1/N_{e',k}} E_k$ where $N_{e,k}$ denotes the number of lower-court cases linked to event e at event time k, and E_k is the number of distinct events observed at horizon k. The normalization term $\sum_{e'} 1/N_{e',k}$ sums over all events e' contributing observations at k.

Figure A.4 plots, for each event time k, the number of equally weighted observations equivalent in information to the weighted sample, defined as $ESS_k = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}$. The effective-sample-size (ESS) distribution is noticeably flatter than the raw count distribution, indicating that weighting prevents a few large judges from disproportionately influencing the estimates.

Throughout the paper, we report unweighted estimates as our main specification, which can be interpreted as the effect of reversals on a randomly drawn lower-court case. For robustness, we report weighted estimates and interpret them as event-level average effects of reversals, following Baker et al. (2025).

Specification and Estimation We estimate the following model for lower-court case i decided by judge j that falls into period k around event e:

$$Y_{ijek} = \sum_{\kappa \in K} [\alpha_k \mathbf{1}(k \in \kappa) + \beta_k \mathbf{1}(k \in \kappa) \cdot \mathbf{1}\left(\text{Reversal}_e = 1\right)] + \mathbf{X}_e \boldsymbol{\phi} + \epsilon_{ijek}, \tag{4.1}$$

where the outcome variable Y_{ijek} is an indicator equaling one if case i is granted in the lower court. $K = \{[-6,-3], \{-2\}, \{0\}, \{1\}, \{2\}, [3,6]\}$ indexes the binned periods. The indicator $\mathbf{1}(k \in \kappa)$ equals one if period k for lower-court case i falls within binned interval $\kappa \in K$. The period before the event ($k = -1$) serves as the reference period. The reversal indicator, $\mathbf{1}(\text{Reversal}_e = 1)$, equals one if event e corresponds to an appeal grant and zero if the appeal is denied. The vector $\mathbf{X}_e$

includes appeal fixed effects, appeal ruling year-by-day of week fixed effects, and appeal ruling month-by-day of week fixed effects. Standard errors are clustered at the judge level. We are primarily intereted in $\hat{\beta}_k$, the estimates of the coefficients on the interaction between the period indicators and the reversal indicator, which trace the dynamic effects of reversals over event time.

As previously discussed, not all reversals provide equally informative feedback to judges. After a surprise reversal, i.e., an unexpected appeal grant following a confident denial, offers a salient signal that the judge's prior assessment was incorrect, triggering meaningful belief updates. To test this learning mechanism, we use the random forest predictions discussed in Section 3.2 as a proxy for judges' subjective beliefs about the probability of appeal grants and estimate the following specification:

$$Y_{ijek} = \sum_{\kappa \in K} \left[a_k \mathbf{1}(k \in \kappa) + b_k \mathbf{1}(k \in \kappa) \cdot \mathbf{1}\left(\text{Surp. Rev.}_e = 1\right) \right] + \mathbf{X}_e \boldsymbol{\varphi} + \varepsilon_{ijek}, \quad (4.2)$$

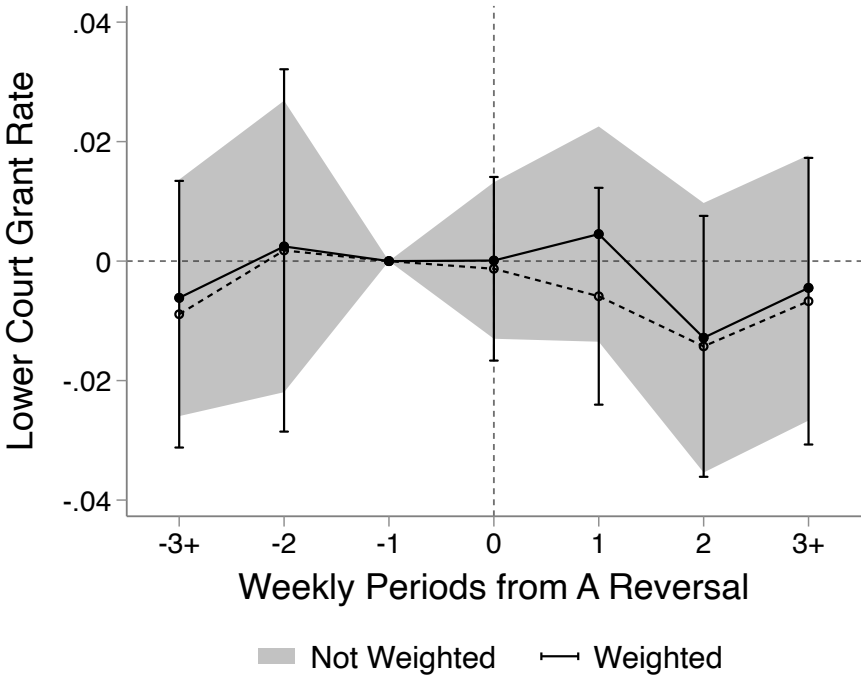
where the surprise reversal indicator $\mathbf{1}(\text{Surp. Rev.}_e = 1)$ equals one if event e corresponds to a surprise reversal (an appeal that is granted despite being predicted to be denied), and zero if the appeal grant was predicted. Again, We are intereted in $\hat{b}_k$, the estimates of the coefficients on the interaction between the period indicators and the surprise reversal indicator. Focusing on surprise reversals allows us to isolate the learning and corrective component of judicial behavior, capturing how judges adjust their decision thresholds after receiving unexpected negative feedback rather than simply responding mechanically to an appeal outcome.

4.2 Event Study Results: Lower Court Grant Rate

Figure 2 presents the event-study estimates of the effect of reversals on lower-court grant rates. We find no evidence of pre-trends, supporting the validity of the empirical design, and no significant post-event changes in grant rates. This suggests that, when examined over a longer horizon, reversals do not have a lasting impact on subsequent lower-court decisions. Figure A.5 reports the unbinned event-study estimates, showing that the results are robust to alternative weighting schemes and binning choices.

In contrast, we observe a clear increase in lower-court grant rates in the weeks

Figure 2: Effect of Reversals on Lower Court Grant Rate

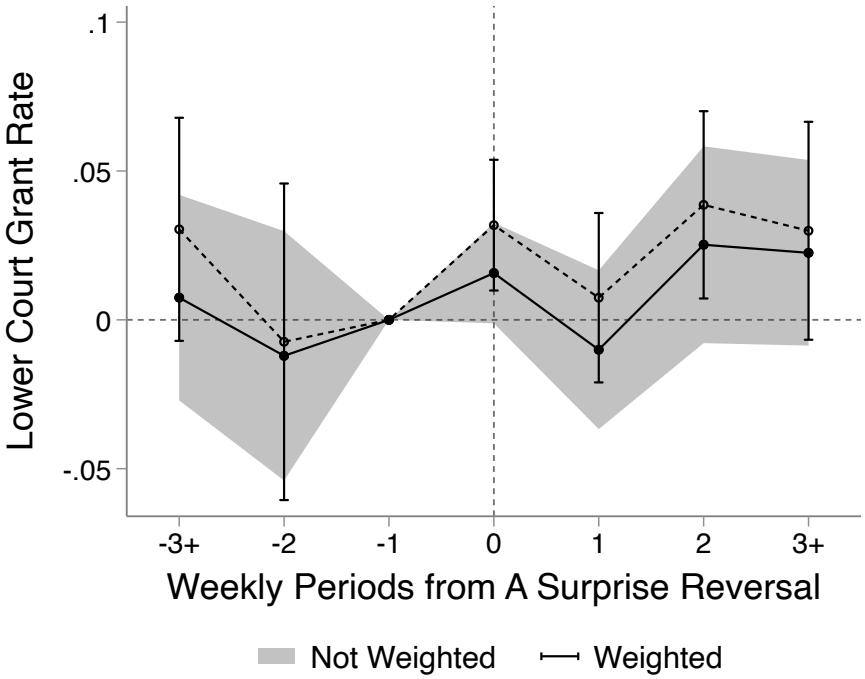


Notes: The figure plots $\hat{\beta}_k$, estimated coefficients of the interaction term of the period indicators and the reversal indicator of Specification (4.1), comparing unweighted and event-equal weighted estimates. The sample includes all appeal decisions from 1986 to 2019 as events, trimmed using the nearest-event attribution method described in Section 4.1. Weights are calculated on the estimation sample as described in Section 4.1. The x-axis indexes seven-day periods around each appeal decision. All estimations include controls for appeal fixed effects and appeal ruling calendar time fixed effects. Standard errors are clustered at the judge level. Unweighted estimates are shown as solid dots with shaded 95% confidence intervals, weighted estimates are shown as open circles with bracketed 95% confidence intervals.

following a surprise reversal, as shown in Figure 3. This effect remains robust when applying the event-equal weighting method. In Figure A.6, we plotted the unbinned estimations, which confirm robustness to the choice of binning. These findings indicate that judges do not respond uniformly to all reversals: following surprise reversals relative to predicted reversals, judges become more lenient. Accordingly, we focus on the effect of surprise reversals and the estimation of Specification (4.2) in what follows. In Figure A.7, we show the unbinned estimation results under the two sample-trimming methods described in Section 4.1. The results remain stable across trimming rules, particularly around the treatment period, indicating that the observed effects are not driven by sample selection.

The results so far suggest that judges react to surprise reversals by increasing

Figure 3: Effect of Surprise Reversals on Lower Court Grant Rate



Notes: The figure plots $\hat{b}_k$, estimated coefficients of the interaction term of the period indicators and the surprise reversal indicator of Specification (4.2), comparing unweighted and event-equal weighted estimates. The sample includes all appeal decisions from 1986 to 2019 as events, trimmed using the nearest-event attribution method described in Section 4.1. Weights are calculated on the estimation sample as described in Section 4.1. The x-axis indexes seven-day periods around each appeal decision. All estimations include controls for appeal fixed effects and appeal ruling calendar time fixed effects. Standard errors are clustered at the judge level. Unweighted estimates are shown as solid dots with shaded 95% confidence intervals, weighted estimates are shown as open circles with bracketed 95% confidence intervals.

their subsequent grant rates. Since surprise reversals can be effectively deemed as negative evaluation signals, this pattern is consistent with judges attempting to reduce their probability of future reversals by granting more cases, thereby sending fewer to the appeal court. However, this behavioral adjustment raises an important question: does the change reflect genuine learning and quality improvement, where overly strict judges update their beliefs to better align with the appeal court, or does it instead represent a mechanical response, where judges become excessively lenient simply to avoid being reversed again? Next, we examine how the diligence of lower court judges adjusts following surprise reversals.

4.3 Event Study Results: Lower Court Hearing Diligence

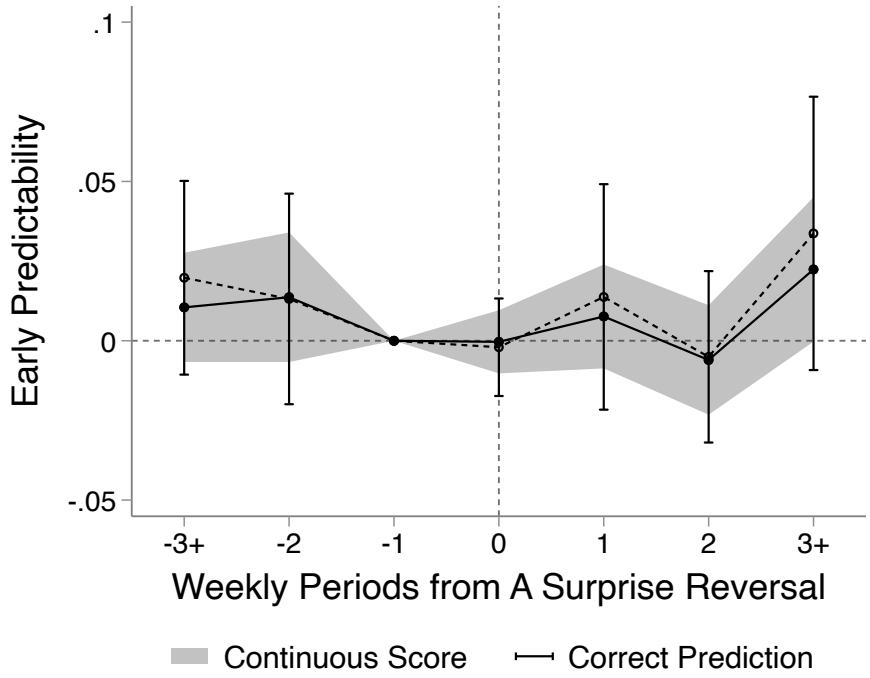
We consider two dimensions of lower court hearing diligence: the outcome dimension, captured by the early predictability of rulings, and the procedural dimension, reflected in the characteristics of the hearing process.

Early predictability of lower court rulings In Section 3.1, we built and tuned the early-prediction models to capture how accurately lower-court decisions can be predicted by random-forest based models using only pre-hearing information as inputs. Here, we employ the best-performing model with judge identifier, applicant nationality, opening year-month, and filing year-month as predictors to generate both binary predictions (equals 1 if granted) and continuous prediction probabilities $\Pr_{ep}(\text{grant})$ (closer to 1 indicates higher likelihood of a grant). We use these early-prediction results to construct two early predictability measures: a binary indicator equal to one if the early-prediction model correctly predicts the actual lower-court decision, and a continuous accuracy measure defined as $1 - |\Pr_{ep}(\text{granted}) - \mathbf{1}(\text{granted})|$, where $\mathbf{1}(\text{granted})$ is the binary indicator equals one if the case is actually granted. This measure ranges from zero (worst performance) to one (perfect prediction) and increases monotonically with predictive accuracy.

We re-estimate Specification (4.2) using the two early-predictability measures and present the results in Figure 4. Relative to the seven-day period immediately preceding a surprise reversal, the early predictability of lower-court decisions does not vary significantly. The results are robust to the choice of outcome variables: both the binary indicator and the continuous measure yield nearly identical estimates. For subsequent analysis, we focus on the continuous measure, which provides more precise estimates.

We observe a mild and statistically insignificant "Ashenfelter's dip" in the periods leading up to surprise reversals. To better understand this pattern, we estimate Specification (4.1) to examine the effect of reversals and present the results in Figure A.8. Similar to our main results, early predictability does not change significantly after a reversal. However, in contrast to Figure 4, we observe an insignificant upward pre-trend: judges' rulings are slightly less early-predictable leading up to a reversal compared with an affirmation. The opposite direction of these pre-trends indicates that the dip observed in Figure 4 is unlikely to reflect judges'

Figure 4: Effect of Surprise Reversals on Lower Court Early Predictability



Notes: The figure plots plots $\hat{b}_k$, estimated coefficients of the interaction term of the period indicators and the surprise reversal indicator of Specification (4.2), with two outcome variables described in Section 4.3: the continuous measure of early predictability and the binary indicator of correct early predictions. The sample includes all appeal decisions from 1986 to 2019 as events, trimmed using the nearest-event attribution method described in Section 4.1. All estimations are unweighted and include controls for appeal fixed effects and appeal ruling calendar-time fixed effects. Standard errors are clustered at the judge level. The x-axis indexes seven-day periods around each appeal decision. Estimates using the continuous early-predictability score are shown as solid dots with shaded 95% confidence intervals, while estimates using the binary indicator for correct early predictions are shown as open circles with bracketed 95% confidence intervals.

anticipation of upcoming surprise reversals. Instead, the evidence is more consistent with mild anticipation effects before predicted reversals, when judges may deviate from routine decision patterns and extract more information from hearings more effectively as they perceive a higher risk of being reversed. Overall, these patterns support our interpretation that surprise reversals constitute unanticipated feedback, whereas predicted reversals are at least partly expected.[11] The results remain robust to alternative binning and weighting schemes as shown in Figure A.9.

[11] In unreported robustness checks, we reestimate the specification in Figure 4 using an indicator equal to one for surprise reversals and zero for affirmations. The results show no pre-trend, confirming that judges do not systematically anticipate surprise reversals.

Diligence in lower court hearings Using the constructed measures of early predictability, we have examined the effect of surprise reversals on how much lower-court judges incorporate information from hearings into their rulings. We next turn to hearing-level data to directly test whether judges alter their hearing diligence after surprise reversals. To do so, we adapt the event-study framework described in Section 4.1: For each judge j on date t, we compute a measure of hearing diligence and assign that observation to one of 13 event-time periods indexed by $k \in [-6, \ldots, 0, \ldots, 6]$. We then estimate:

$$Y_{jtek} = \sum_{\kappa \in K} \left[\theta_k \mathbf{1}(k \in \kappa) + \delta_k \mathbf{1}(k \in \kappa) \cdot \mathbf{1}\left(\text{Surp. Rev.}_e = 1\right)\right] + \mathbf{X}_e \boldsymbol{\varphi} + \varepsilon_{ijek}, \quad (4.3)$$

where the regressors are defined as in Specification (4.2), and the outcome variable Y_{jtek} denotes judge j's hearing diligence on date t that falls in period k of event e.

For each proceeding, we use several key variables from the hearing-level data to construct our measures of hearing diligence. Specifically, we count the total number of hearings, compute the average duration of adjournments, and calculate both the share of adjournments granted for preparation and the share of hearings held on individual calendars. Table A.6 reports summary statistics for these measures from 1984 to 2019 across all proceedings.[12] Among proceedings linked to our main sample, we observe an average of about 3.8 hearings per case, an average adjournment duration of roughly 139 minutes, 8% of adjournments granted for preparation, and 35% of hearings held on individual calendars. To align these measures with our event-study framework, we compute the four measures for each proceeding handled on date t by judge j, and then average across all proceedings within each judge-day unit.

We present the unweighted estimation results of Specification 4.3 in Table 1. The number of hearings per proceeding shows a small but statistically significant increase in the week right after a surprise reversal, suggesting that judges may become more diligent by scheduling additional hearings before issuing subsequent rulings after strong negative evaluation signals. This result is robust to alternative weighting schemes (see Figure A.10) and remains persistent, though statistically insignificant, in later periods. We find no evidence of significant adjust-

[12] We winsorize hearing duration at the top 1% of the full hearing-level distribution to mitigate extreme right skewness before constructing any measures.

ments in the composition of hearings: the shares of preparation adjournments and individual-calendar hearings remain largely stable following a surprise reversal. Finally, adjournment duration exhibits a mild U-shaped pattern around surprise reversals, echoing our earlier findings on the effects of surprise reversals on early predictability. Given the presence of a noticeable pre-trend, we interpret these results as descriptive patterns rather than causal effects.

In the hearing-level data, we observe hearings presided over by judges not only for cases in the main asylum sample but also for other proceedings including non-asylum, procedural, and administrative cases, which together account for roughly 81% of all hearings. To examine whether surprise reversals in asylum cases affect judges' broader hearing behavior, we reconstruct the four diligence measures using all hearings, not just those linked to asylum cases, and re-estimate Specification 4.3. The results, reported in Table A.6, show no evidence that surprise reversals in asylum cases influence judges' general hearing diligence in terms of the number of hearings per proceeding, the average duration of adjournments, or the share of preparation adjournments. We do, however, find a marginally significant and persistent increase in the share of individual-calendar hearings following surprise reversals, suggesting a modest reallocation of effort across hearing formats.

Taken together, our results suggest that judges' hearing diligence does change after surprise reversals: Following unexpected negative feedback, judges' rulings become less early predictable, the average number of hearings per proceeding increases, and the other measures of hearing diligence do not vary significantly. Although these effects are marginally significant and should be interpreted with caution, they provide a consistent narrative: judges respond to surprise reversals not by lowering their hearing diligence. Combined with the observed increase in grant rates after surprise reversals, this pattern suggests that such singals are likely to correct overly strict or systematically biased decision-making.

5 Attention and Learning of Lower Court Judges

The empirical evidence presented so far shows that judges react to surprise reversals and adjust their ruling behavior following these unexpected evaluation signals. However, these average effects mask substantial heterogeneity across judges. In this section, we delve deeper into the learning mechanisms of lower-

Table 1: Effect of Surprise Reversals on Lower Court Asylum Hearing Diligence

	(1) No. hearings	(2) Avg. duration	(3) Prep. Share	(4) Ind. Share
$\mathbf{1}(k \in [-6,-3])$	-0.007	-5.458***	0.011	0.004
	(0.010)	(2.028)	(0.011)	(0.016)
$\mathbf{1}(k=-2)$	-0.012	-4.553**	0.009	-0.024**
	(0.014)	(2.156)	(0.011)	(0.012)
$\mathbf{1}(k=0)$	-0.025**	-2.280	0.012	-0.019*
	(0.012)	(1.759)	(0.009)	(0.011)
$\mathbf{1}(k=1)$	-0.010	-0.803	0.001	-0.0162
	(0.012)	(1.880)	(0.009)	(0.011)
$\mathbf{1}(k=2)$	-0.023*	-4.604*	0.006	-0.002
	(0.013)	(2.426)	(0.011)	(0.011)
$\mathbf{1}(k \in [3,6])$	-0.013	-4.241*	0.012	-0.016
	(0.015)	(2.241)	(0.010)	(0.013)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k \in [-6,-3])$	0.010	6.126***	-0.007	-0.011
	(0.010)	(1.972)	(0.011)	(0.014)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k=-2)$	0.010	3.731*	-0.008	0.014
	(0.013)	(2.003)	(0.009)	(0.011)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k=0)$	0.017*	1.881	-0.009	0.013
	(0.010)	(1.564)	(0.008)	(0.010)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k=1)$	0.005	1.386	0.001	0.013
	(0.010)	(1.671)	(0.008)	(0.010)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k=2)$	0.018	4.475**	-0.007	0.007
	(0.012)	(2.232)	(0.010)	(0.010)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k \in [3,6])$	0.009	3.521	-0.013	0.014
	(0.013)	(2.166)	(0.009)	(0.012)
adj. R^2	0.151	0.367	0.133	0.117
N	911,110	659,383	621,481	701,833

Note: This table reports $\hat{\theta}_k$ and $\hat{\delta}_k$, estimated coefficients of the period indicators and the interaction term of the period indicators and the surprise reversal indicator of Specification (4.2), respectively. Each column presents results using one of the four proceeding-level hearing diligence measures constructed in Section 4.3 on asylum-related hearings: the number of hearings, the average duration of adjournments, the share of preparation adjournments, and the share of individual calendar hearings. The sample includes all appeal decisions from 1986 to 2019 as events, trimmed using the nearest-event attribution method described in Section 4.1. All estimations are unweighted and include controls for appeal fixed effects and appeal ruling calendar-time fixed effects. Standard errors clustered at the judge level are reported in parentheses. Asterisks denote significance at the 1% (***), 5% (**), and 10% (*) levels.

court judges in response to feedback from the appeal court. Specifically, we ask which types of judges are most attentive to signals in appeal decisions, and what

they appear to learn from such signals. To organize our analysis, we first present a simple model of lower court reversals that links judges' learning and attention to the observed reversal rate. We then bring the model to the data using both reduced-form diagnostics and marginal treatment effects (MTEs) that trace how reversal probabilities respond to marginal changes in latent grant propensity for attentive and inattentive judges.

5.1 A Simple Model of Reversals

Each case i has a latent quality $q_i \in [0,1]$ that summarizes how deserving it is of asylum from the appeal court's perspective. The distribution of q_i has density $f(q) > 0$ and CDF $F(q)$, with f and F continuously differentiable. A lower court judge j with leniency level l_j applies a threshold $c_j = c(l_j)$, where the mapping $c(\cdot)$ is differentiable with $c'(l) < 0$. Judge j grants asylum if and only if

$$D_i := \mathbf{1}\{q_i \geq c_j\} = 1.$$

An applicant denied in the lower court may file an appeal. We capture this with a simple appeal rule: an appeal is filed if and only if

$$A_i := \mathbf{1}\{q_i \geq a\} = 1,$$

where $a < c^*$ is the appeal-benefit threshold.[13] In other words, among denied cases, only those with sufficiently high quality find it worthwhile to appeal. This representation nests richer models in which the benefit of appeal is increasing in q_i. The appeal court applies a fixed standard $c^* \in (0,1)$: an appeal is granted if and only if $q_i \geq c^*$. We focus on the empirically relevant case where judges are stricter than the appeal board, i.e. $c_j > c^*$.

A reversal occurs when a case is denied in the lower court but granted on appeal. Let

$$Y_i = \mathbf{1}\{q_i < c_j,\ q_i \geq c^*\}$$

[13] Equivalently, one can think of a cost of appeal that is decreasing in q; this is simply the cutoff representation.

denote a reversal for case i. The reversal rate for judge j with cutoff c_j is then

$$R(c_j) := \Pr(Y_i = 1 \mid c_j) = \int_{c^*}^{c_j} f(q)\,\mathrm{d}q.$$

The reversal rate can be decomposed into a product of three components:

$$\begin{aligned} R(c_j) &= \Pr(Y_i = 1 \mid c_j) \\ &= \Pr(D_i = 0 \mid c_j) \cdot \Pr(A_i = 1 \mid D_i = 0, c_j) \cdot \Pr(S_i = 1 \mid A_i = 1, D_i = 0, c_j), \end{aligned}$$

where $D_i = 0$ indicates a denial in the lower court, $A_i = 1$ indicates that a denied case proceeds to appeal, and $S_i = 1$ indicates that an appealed case is granted by the appeal court. We refer to the appeal rate of all denied cases $E(c_j) := \Pr(A_i = 1 \mid D_i = 0, c_j)$ as the selection (*extensive*) margin, and the appeal grant rate conditional on appealing $I(c_j) := \Pr(S_i = 1 \mid A_i = 1, D_i = 0, c_j)$ as the success (*intensive*) margin. Under the appeal rule described above, denials occur for $q_i < c_j$, and among them, appeals are filed when $q_i \geq a$. Then we can rewrite the two margins as

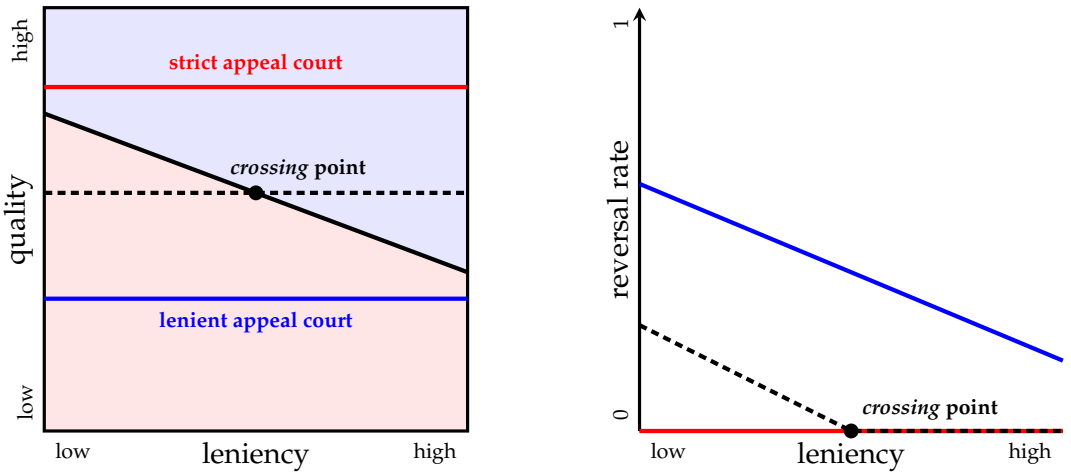
$$E(c_j) = \frac{F(c_j) - F(a)}{F(c_j)}, \qquad I(c_j) = \frac{F(c_j) - F(c^*)}{F(c_j) - F(a)}.$$

To build intuition for how reversals behave as a function of judge leniency, Figure 5 provides a stylized illustration of the mapping from case quality and judge leniency to lower-court rulings under three scenarios: a *strict* appeal court with $c^* > \max\{c_j\}$, a lenient one with $c^* < \min\{c_j\}$, and a *crossing* one with $c^* \in [\min\{c_j\}, \max\{c_j\}]$. Figure 5a shows the lower court decision rules, and Figure 5b shows the corresponding reversal–leniency relationship, which Proposition 5.1 formalizes:

Proposition 5.1. *For judges with $c_j > c^*$ and appeal-benefit threshold $a < c^*$, the reversal rate $R(c_j)$ is decreasing in l_j, and both the selection margin $E(c_j)$ and the success margin $I(c_j)$ are also decreasing in l_j.*

Intuitively, stricter judges (higher c_j, lower leniency) deny more cases that the appeal court would grant, generating more reversals overall. Among denials, they create a large pool of strong cases with both high appeal rates and high success rates on appeal. As judges become more lenient (lower c_j), these strong denials are granted at first instance and disappear from the denial pool, causing the reversal

Figure 5: A Stylized Illustration of Leniency and Reversal



(a) Lower Court Ruling (b) Reversals

Notes: Panel 5a illustrates the stylized relationship case quality, judge leniency and lower court decisions across appeal courts of varying stringency. Panel 5b displays the corresponding appeal grant rate by leniency for appeal courts of corresponding stringency. Coordinates and slopes are schematic and chosen for expositional clarity.

rate and both the selection and success margins to decline in leniency. Detailed proofs are provided in Section C.

So far, thresholds c_j are static. We now model how attentive judges learn from reversals and correct their threshold bias dynamically. Let $c_{j,t}$ be judge j's threshold in period t. After observing reversals, the judge compares their own threshold to the appeal board's standard c^* and updates according to

$$c_{j,t+1} = c_{j,t} - \lambda_j \left(c_{j,t} - c^* \right),$$

where $\lambda_j \in [0, 1]$ is an attention parameter: for inattentive judges, $\lambda_j = 0$ and they never update; for attentive judges, $\lambda_j \in (0, 1)$ and they partially adjust toward the appeal court's cutoff. Solving this simple recursion, the judge's deviation from the appeal court threshold follows

$$c_{j,t} - c^* = (1 - \lambda_j)^t \left(c_{j,0} - c^* \right).$$

The next proposition formalizes how only attentive judges learn to correct their threshold bias over time:

Proposition 5.2. *For an attentive judge with $\lambda_j > 0$, the threshold converges to the appeal*

court threshold, $\lim_{t\to\infty} c_{j,t} = c^*$*. For an inattentive judge (*$\lambda_j = 0$*), the threshold remains fixed at its initial value,* $c_{j,t} = c_{j,0} > c^*$ *for all* t*.*

The proposition formalizes the idea that an attentive judge gradually moves his threshold toward the appeal court standard c^*, while an inattentive judge keeps using his initial, stricter cutoff $c_{j,0}$. The proof is straightforward.

To understand the empirical implications of this divergence in thresholds, we examine how sensitive reversals and their components are to marginal changes in the threshold itself. The relative sensitivities at c^* versus $c_{j,0}$ will play a key role in interpreting the variation we observe across the two types of judges. Define the local sensitivity functions of the reversal rate and the two margins with respect to thresholds as

$$\psi_R(c) = \frac{\partial R}{\partial c}(c) = f(c),$$
$$\psi_E(c) = \frac{\partial E}{\partial c}(c) = \frac{F(a)f(c)}{F(c)^2}, \qquad \psi_I(c) = \frac{\partial I}{\partial c}(c) = \frac{f(c)\left[F(c^*) - F(a)\right]}{\left[F(c) - F(a)\right]^2}.$$

Under regularity conditions on the case-quality distribution, we then have the following lemma:

Lemma 5.1. *Assume that* $f'(c) > 0$ *and* $f'(c)F(c) > 2f(c)^2$ *on* $[c^*, c_{j,0}]$*. Then*

$$\psi_R(c_{j,0}) > \psi_R(c^*), \qquad \psi_E(c_{j,0}) > \psi_E(c^*).$$

If, in addition, $f'(c)\left[F(c) - F(a)\right] > 2f(c)^2$ *on* $[c^*, c_{j,0}]$*,*

$$\psi_I(c_{j,0}) > \psi_I(c^*).$$

The lemma captures a central administrative feature of asylum adjudication. Stricter judges deny many cases that lie just below their cutoff, including a large number that would be granted by the appeal court. When the threshold is high, this "borderline" region contains substantial probability mass, so a small reduction in the cutoff reclassifies many cases. Reversals, appeal rates among denials, and appeal success among appealed denials therefore become highly sensitive to marginal changes in the threshold. In contrast, when a judge's threshold converges to the appeal court standard c^*, the borderline region shrinks drastically: almost all strong cases are correctly granted in the lower court. Consequently, reversals and both margins become much less sensitive to the threshold.

Lemma 5.1 indicates that an inattentive judge's threshold $c_{j,0} > c^*$ remains in this high-sensitivity region, while an attentive judge eventually moves to the low-sensitivity region around c^*. This difference in the local responsiveness of outcomes to the threshold provides the theoretical foundation for the empirical divergence we examine in the next subsections. We prove Lemma 5.1 and discuss the necessary regularity conditions in Section C.

5.2 Characterizing Judge Heterogeneity

Understanding how judges differ in their behavior is central to our analysis of learning. The model points to two key dimensions of judge heterogeneity: judges' attentiveness to feedback signals and their leniency in ruling. In this section, we construct the empirical measures of these two judicial characteristics.

5.2.1 Attentiveness to Feedback

We modify Specification (4.2) to estimate the attentiveness of judges to surprise reversals. Instead of $K = \{[-6,-3],\{-2\},\{0\},\{1\},\{2\},[3,6]\}$, we use an alternative event-time binning $K' = \{[-6,-3],\{0,1\},[2,6]\}$ and estimate for each judge j

$$\begin{aligned} Y_{iek} =& \lambda_1 \mathbf{1}(k \in [-6,-3]) + \mu_1 \mathbf{1}(k \in [-6,-3]) \cdot \mathbf{1}\left(\text{Surp. Rev.}_e = 1\right) \\ &+ \lambda_2 \mathbf{1}(k \in \{0,1\}) + \mu_2 \mathbf{1}(k \in \{0,1\}) \cdot \mathbf{1}\left(\text{Surp. Rev.}_e = 1\right) \\ &+ \lambda_3 \mathbf{1}(k \in [2,6]) + \mu_3 \mathbf{1}(k \in [2,6]) \cdot \mathbf{1}\left(\text{Surp. Rev.}_e = 1\right) \\ &+ \mathbf{X}_e \boldsymbol{\psi} + \nu_{iek}, \end{aligned} \tag{5.1}$$

where the two 7-day periods immediately before the reversal serve as the reference category, while controlling for long leads (at least three 7-day periods before) and long lags (at least two 7-day periods after). We use the estimated coefficient $\hat{\mu}_2$ to construct our Measures of attentiveness, capturing each judge's reaction in the outcome variable during the first two periods following a surprise reversal. This binning strategy provides tighter control for pre-trends, reduces noise in judge-level estimation, and ensures that the estimated effects reflect sustained behavioral adjustments rather than mechanical or transient responses.

Grant rate attentiveness Figure A.11 plots the estimated effects of surprise reversals on subsequent rulings for 354 judges with valid estimates out of 698 judges in the full sample. Both the median and mean of these estimates are positive, consistent with the aggregate event-study results above: on average, judges increase their grant rates following a surprise reversal. At the same time, judges' response adjustments in granting probabilities following negative evaluation signals are not unidirectional and vary widely across individuals. Judges who are initially overly strict could increase their probability of granting asylum after a surprise reversal, whereas judges who are overly lenient tend to decrease it. This could be due to a bias-correction mechanism, where feedback from the appeal court helps judges recalibrate their over-strict or over-lenient ruling tendencies. Accordingly, we take the absolute value of the estimated effects as our main attentiveness measure.

Hearing diligence attentiveness The second dimension of attentiveness we consider speaks to the hearing diligence of judges. We construct two attentiveness measures to capture judges' behavioral adjustments in hearing diligence after surprise reversals.

First, we consider changes in the quality of rulings, proxied by the early predictability of a decision. We construct a judge-level early predictability reaction measure by re-estimating Specification (5.1) with the continuous early predictability score described in Section 4.3 as the outcome variable. Figure A.12 plots the estimated effects of surprise reversals on early predictability for 380 judges with valid estimates. Unlike in the grant-rate dimension, if judges respond to appeal feedback and correct their ruling bias in any meaningful way, we would expect a decline in early predictability following surprise reversals. Accordingly, we use the raw estimated effects as our early predictability reaction measure.

Second, we directly measure changes in the average number of hearings per proceeding. We adopt the same modification used in Specification (5.1) to Specification (4.3) and estimate, for judge j,

$$\begin{aligned} Y_{tek} =& \lambda_1 \mathbf{1}(k \in [-6,-3]) + \mu_1 \mathbf{1}(k \in [-6,-3]) \cdot \mathbf{1}\left(\text{Surp. Rev.}_e = 1\right) \\ &+ \lambda_2 \mathbf{1}(k \in \{0,1\}) + \mu_2 \mathbf{1}(k \in \{0,1\}) \cdot \mathbf{1}\left(\text{Surp. Rev.}_e = 1\right) \\ &+ \lambda_3 \mathbf{1}(k \in [2,6]) + \mu_3 \mathbf{1}(k \in [2,6]) \cdot \mathbf{1}\left(\text{Surp. Rev.}_e = 1\right) \\ &+ \mathbf{X}_e \boldsymbol{\psi} + \nu_{tek}. \end{aligned} \tag{5.2}$$

We then plot the estimates $\hat{\mu}_2$ for 380 judges with valid estimates in Figure A.13. The number of hearings per proceeding is expected to be positively related to the quality of rulings. Therefore, a positive estimate implies that judges become more diligent after surprise reversals. As with the early-predictability reaction measure, we use the raw estimated effects as our hearing-diligence reaction measure.

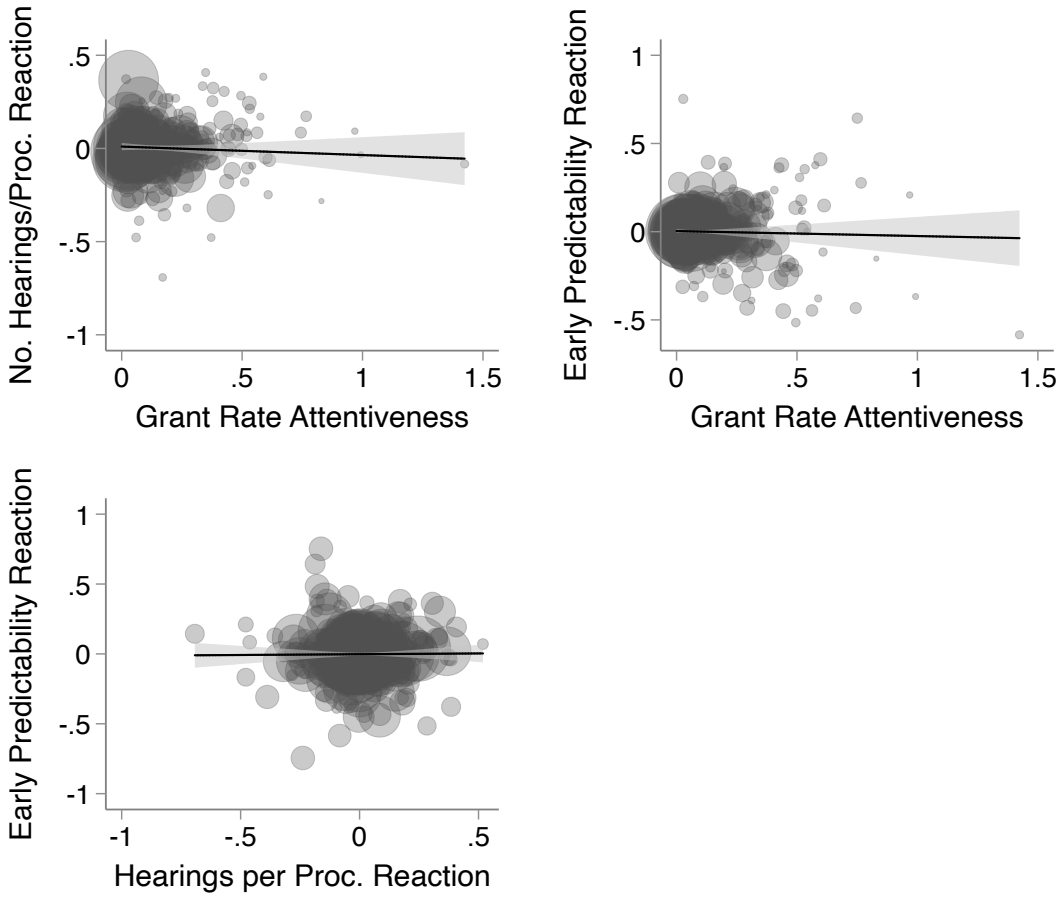
Figure 6 plots the pairwise correlations among the three attentiveness measures. We do not observe significant correlations between any two dimensions. This suggests that the three measures capture distinct aspects of judges' behavioral responses to appeal feedback.

Due to sample size limitations, our main analysis constructs the attentiveness measures at the judge level. Admittedly, judges' attentiveness may vary over time. As a robustness check, we construct the three attentiveness measures for each judge-year unit. Figure A.14 plots the pairwise correlations among the three judge-year level measures and confirms the consistency of our earlier findings: the three measures remain largely orthogonal to one another.

Patterns of judicial attentiveness After constructing the three measures of attentiveness, we examine how they vary with judges' biographical characteristics and over time. Figure 7 shows that, at the judge level, grant-rate attentiveness is significantly and positively associated with the number of years of professional experiences between law school graduation and appointment, as well as with prior non-INS government or academic experiences. These relationships are robust across both univariate and multivariate specifications. In contrast, prior military and NGO experience are marginally associated with lower grant-rate attentiveness, but only in the multivariate specification. In Figure A.15, we show that prior NGO experience is associated with lowering early predictability, and this relationship is robust across specifications. Both military and NGO experience are associated with reducing hearing diligence, but only in the multivariate specification.

We repeat this analysis using the judge-year level attentiveness measures and report the results in Figures A.16 and A.17. INS experience, non-INS government experience, and being male are associated with higher grant rate attentiveness, whereas prior judicial experience and private-practice experience are negatively related to grant rate attentiveness. INS experience and years of judicial service are also negatively correlated with early-predictability reactions, while non-INS gov-

Figure 6: Correlations of Judge Level Attentiveness Measures

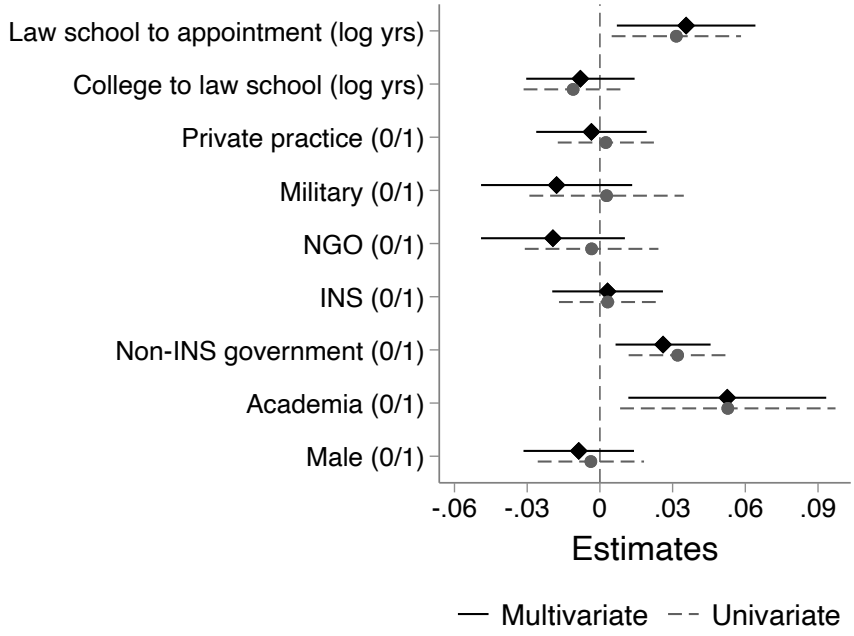


Notes: The figure plots the correlations between the three attentiveness measures described in Section 5.2.1. Each dot represents a judge's attentiveness in the two dimensions shown on the axes. The size of each dot reflects the number of observations used to generate the estimate for the dimension on the x-axis. The fitted line is weighted by the number of observations used to construct each corresponding estimate, and the gray shaded area indicates the 95% confidence interval around the fitted line.

ernment experience and total years of pre-appointment professional experience are associated with an increase in hearings per proceeding. The discrepancies between the judge-level and judge-year level results may partly reflect sample composition changes across years.

Lastly, we examine how the three attentiveness measures evolve over time. Specifically, we compare judge year level attentiveness across six presidential cycles from 1993 to 2016 with that observed during the Trump presidency, controlling for judge fixed effects. Figures A.18, A.19, and A.20 show that all three attentive-

Figure 7: Judge Level Grant Rate Attentiveness and Judge Characteristics



Notes: The figure plots the correlations between the judge-level grant-rate attentiveness measure and judges' biographical characteristics. Diamonds represent coefficients from multivariate regressions including all characteristics as regressors, with solid lines denoting 95% confidence intervals. Circles represent coefficients from univariate regressions, with dashed lines denoting 95% confidence intervals. All regressions are weighted by the number of proceedings ruled by each judge. Standard errors are robust to heteroskedasticity.

ness measures remain relatively stable over time, suggesting that judges' learning patterns are not strongly influenced by changes in the political climate. However, we observe that judges during Democratic administrations (Clinton 1993-2000 and Obama 2009-2016) appear marginally and insignificantly less attentive across all three measures compared with the subsequent Republican administrations (W. Bush 2001-2008 and Trump 2017-2019).

5.2.2 Judge Leniency

Next, we follow Dobbie, Goldin and Yang (2018) and Arnold, Dobbie and Yang (2018) and construct a residualized, leave-out judge leniency measure that accounts for the case-assignment process. Specifically, we first regress lower-court decisions on court-by-year-by-day of week and court-by-month-by-day of week fixed effects, the levels at which cases are quasi-randomly assigned to judges. We

then use the residuals from this regression to calculate the leave-out mean grant rate for each proceeding. In all subsequent analyses, we reconstruct the judge leniency measure within each relevant subsample.

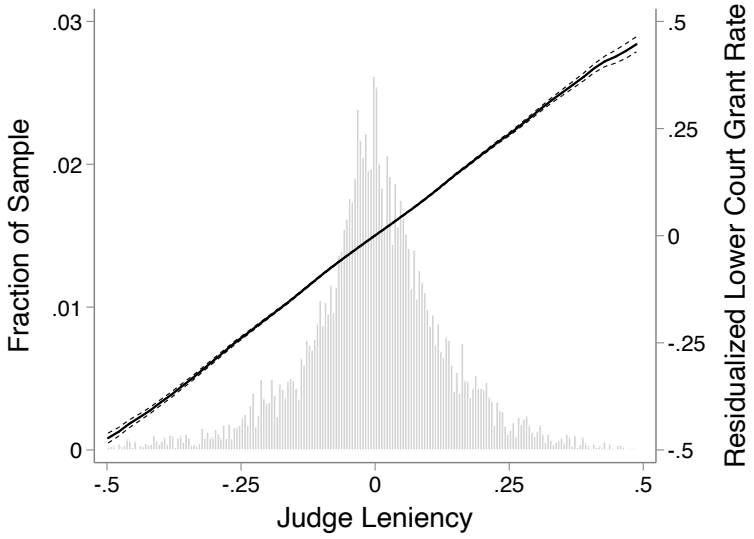
Existence of first stage To validate our leniency instrument, we first document a strong first-stage relationship between lower court grant rates and the constructed leniency measure. We residualize the indicator for lower-court grants using the same set of court-by-time fixed effects and plot a local linear regression of the residualized grant rate against judge leniency. As shown in Figure 8, the residualized grant rate increases monotonically with the leniency measure, confirming a strong and consistent first-stage relationship.

Exclusive restriction Next, we test whether case assignment is random after controlling for the full set of court-by-time fixed effects. Table A.8 reports the results of this randomization test. In Column (1), we find that applicants' origin regions, the presence of a lawyer, the issuance of written decisions, the timing and order of cases within a day, and the number of charges and applications for a case are all significant predictors of lower-court grants. In Column (2), we examine whether these case characteristics are systematically correlated with the constructed judge leniency measure. While leniency remains correlated with applicants' origin regions, we find no significant differences in administrative complexity (number of charges and applications) or assignment procedures (order and timing) across judges with different levels of leniency. We argue that, similar to the role of defendants' race in earlier work (Arnold, Dobbie and Yang, 2018; Dobbie, Goldin and Yang, 2018), the constructed judge leniency measure may not fully capture potential biases associated with these underlying factors. To the extent that the exclusion restriction is violated, the results we present can be interpreted as the causal impact of being assigned to a more or less lenient lower court judge.

Monotonicity The final assumption we check is the monotonicity, which requires that a case granted by a strict judge would also be granted by lenient judges, and a case denied by a lenient judge would also be denied by strict judges. This assumption is necessary for estimating marginal treatment effects using the standard local instrumental variables (LIV) approach (Angrist, Imbens and Rubin, 1996;

Heckman and Vytlacil, 2005). A testable implication of monotonicity is that the first-stage relationship between judge leniency and grant rates should be positive across all subsamples. We assess this by plotting the residualized judge leniency measures calculated for different subsamples defined by applicants' origins and case characteristics in Figure A.21. Each subplot reports the corresponding correlation coefficient and standard error. Consistent with the monotonicity assumption, we find that judge leniency measures across subsamples are always significantly positively correlated, suggesting that judges' relative tendencies are stable across observably different cases.

Figure 8: Judge Leniency and Lower Court Grant Rates



Notes: The figures plot the relationships between judge leniency and lower-court grant rate. The leave-out judge leniency measure is constructed as described in Section 5.2.2. The solid line shows a local linear regression of residualized lower-court asylum grants on judge leniency. Lower court grants are residualized using the same set of court-by-time fixed effects applied in constructing the leniency measure in the full sample. Standard errors are clustered at the judge level, and dashed lines denote 95% confidence intervals.

5.3 Testing Model Predictions

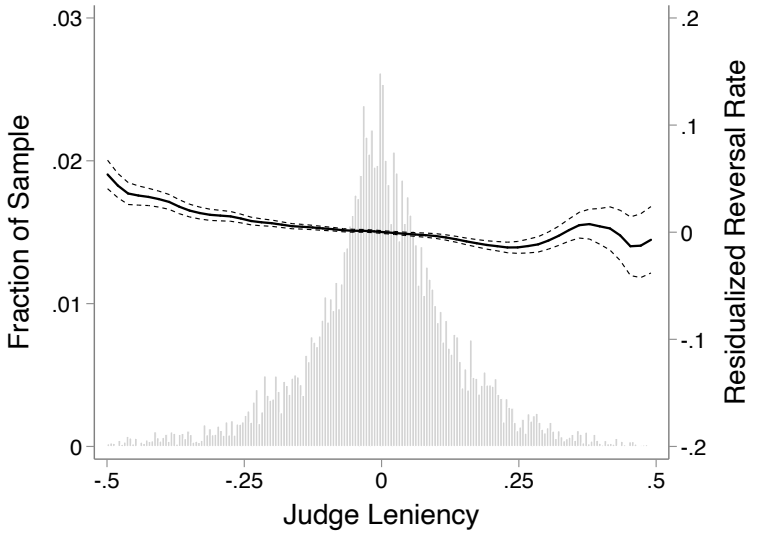
After establishing the two dimensions of judicial heterogeneity, we present the reduce-form evidence and marginal treatment effect (MTE) estimation results to test the predictions of our model.

5.3.1 Reduced-Form Evidence

The model in Section 5.1 delivers a set of clear reduced-form predictions. Proposition 5.1 predicts that when judges employ thresholds stricter than the appeal court standard c^*, the reversal rate and its extensive and intensive margins should all be decreasing in leniency. Moreover, the slope of this relationship should be steeper in the strict region, where judges operate far from c^*, and gradually flatten as thresholds approach c^*.

In Figure 9, we plot a local linear regression of residualized lower-court reversals on judge leniency. Consistent with the model, we find a significant negative correlation: more lenient judges face lower reversal risk up to a leniency threshold, beyond which the residualized reversal rate stabilizes around zero. This flattening is consistent with the implication that once thresholds approach the appeal court standard, further changes in leniency have limited impact on reversals.

Figure 9: Judge Leniency and Lower Court Reversal Rates



Notes: The figure plots the relationship between judge leniency and the lower-court reversal rate. The leave-out judge leniency measure is constructed as described in Section 5.2.2. The solid line shows a local linear regression of residualized lower-court reversals, defined as lower-court denials that are granted in the appeal court, on judge leniency. Lower-court reversals are residualized using the same set of court-by-time fixed effects applied in constructing the leniency measure in the full sample. Standard errors are clustered at the judge level, and dashed lines denote 95% confidence intervals.

Next, we test the predictions for both margins, appeal among denials and appeal success among appealed denials, in Figure A.22. Figure A.22a plots the residualized selection (extensive) margin against the constructed leave-out leniency. We

find a statistically significant negative correlation: cases denied by more lenient judges are indeed less likely to be appealed. Figure A.22b plots the residualized success (intensive) margin against leniency, and again we find a significant negative correlation: among denials that are appealed, those handled by more lenient judges are less likely to be granted on appeal.

A key assumption of the model is that the appeal court standard c^* is independent of lower-court judges' leniency. In Figure A.23, we test this assumption. Using the random-forest model trained on realized appeal decisions described in Section 3.2, we generate predicted probabilities of appeal grants for all cases as a proxy for appeal-court leniency. We then residualize these predicted probabilities and regress them on judge leniency. The results reveal a modest positive relationship: cases handled by more lenient judges are slightly more likely to be granted by the appeal court. In Figure A.24, we calibrate our stylized illustration to reflect this deviation and show that it modestly flattens the reduced-form reversal–leniency relationship, but does not alter the key comparative static that reversals are decreasing in leniency and steepest in the strict region.

Proposition 5.2 and Lemma 5.1 predict that attentive judges converge to the appeal-court standard while inattentive judges do not. This implies that reversal rates should be less sensitive to leniency for attentive judges than for inattentive judges. We use the constructed attentiveness measures to classify proceedings according to the attentiveness of the presiding judge and test these predictions. Specifically, within each ruling year, we rank proceedings by the attentiveness values of their ruling judges. For our main measure, grant-rate attentiveness, we define the top tercile of proceedings as those ruled by *attentive* judges and the bottom tercile as those ruled by *inattentive* judges. For the two hearing-diligence attentiveness measures, we classify proceedings based on the sign of the corresponding judge's attentiveness measure, labeling judges that decrease early predictability or increase the daily average number of hearings per proceeding as *attentive*, and others as *inattentive*.[14] We then replicate the analysis presented in Figures 8 and 9 for

[14]These different classification rules reflect the distinct nature of the three attentiveness measures. The grant-rate attentiveness measure is defined as the absolute value of the estimated effect of surprise reversals on lower-court grants. As shown in Figure A.11, the distribution of the raw estimates is concentrated around zero; hence, splitting the absolute values into terciles distinguishes inattentive judges (values near zero) from attentive judges (values significantly different from zero). For the hearing-diligence attentiveness measures, we rely on the raw estimates of the effects of surprise reversals, since positive and negative reactions have clear monotonic interpretations. In the

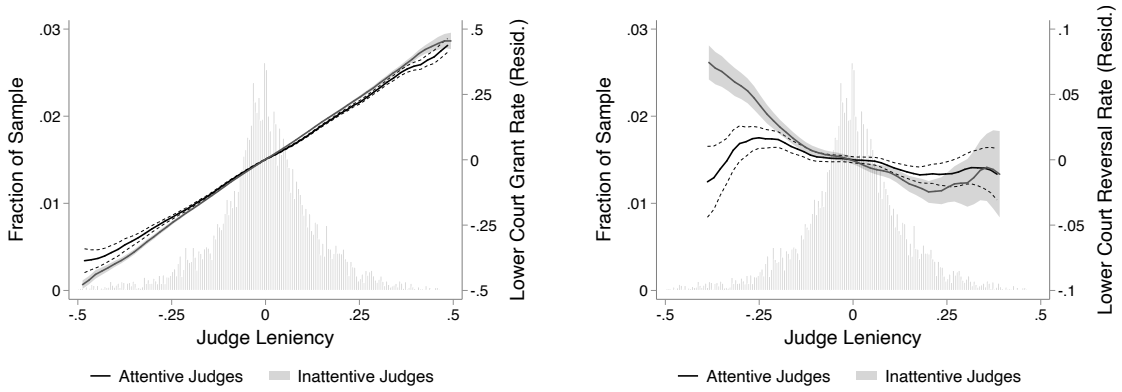
proceedings ruled by attentive and inattentive judges separately.

Grant-rate attentiveness In Figure 10, we compare proceedings ruled by attentive and inattentive judges. For both groups, the residualized lower-court grant rate is strongly and significantly positively correlated with judge leniency. Importantly, the slope is significantly flatter for proceedings ruled by attentive judges than for those ruled by inattentive judges. This flatter correlation is consistent with attentive judges aligning their rulings more closely with the appeal-court standard, as suggested by Figure A.23. When examining the relationship between the residualized reversal rate and judge leniency, we find a similarly flatter slope for proceedings ruled by attentive judges, indicating that reversal risk is much less sensitive to individual leniency for attentive judges. In Figure A.25, we show that the patterns observed in Figure 10 are robust and even more pronounced when using the judge-year level construction of attentiveness measures. The disparity between attentive and inattentive subsamples also remains robust under an alternative classification based on top-versus-bottom halves, as shown in Figure A.26. Taken together, these results validate the model's prediction that attentive judges learn from negative evaluation signals and converge toward the appeal-court standard, whereas inattentive judges do not.

Hearing-diligence attentiveness We then repeat the same exercise by splitting the sample using the two hearing-diligence attentiveness measures. First, we classify proceedings based on whether the ruling judges' early predictability decreases after surprise reversals. As discussed earlier, if judges reduce early predictability after surprise reversals, they are likely incorporating more information during the hearing process and thus making higher-quality rulings. Accordingly, judges with negative reactions are considered more attentive. Figure A.27a shows that the two groups of judges do not differ significantly in the correlation between the residualized lower-court grant rate and leniency. Figure A.27b shows that, among judges who reduce early predictability after surprise reversals, the correlation between residualized reversal risk and leniency is more modest than for positive-reacting judges, although the difference is less pronounced than in the grant-rate attentiveness results.

Appendix, we report the robustness of our results to alternative classification methods.

Figure 10: Leniency and Lower Court Outcomes by Grant Rate Attentiveness



(a) Grant Rate (b) Reversal Rate

Notes: The figures plot the relationships between judge leniency and lower-court outcomes for attentive and inattentive judges, defined by the magnitude of the effect of surprise reversals on grant rates and classified into the top and bottom terciles, as described in Section 5.3. The leave-out judge leniency measure is constructed as described in Section 5.2.2 for the two subsamples. The left panel shows local linear regressions of residualized lower-court asylum grants on judge leniency, and the right panel shows local linear regressions of residualized lower-court reversals, defined as lower-court denials that are granted on appeal on judge leniency. Both lower-court outcomes are residualized using the same set of court-by-time fixed effects applied in constructing the leniency measure. In both panels, the black line represents the subsample of cases ruled by attentive judges, with dashed lines indicating 95% confidence intervals, while the gray line represents the subsample of cases ruled by inattentive judges, with shaded areas denoting 95% confidence intervals. Standard errors are clustered at the judge level.

Next, we classify proceedings based on whether the ruling judges increase the average number of hearings per proceeding after surprise reversals. We hypothesize that ruling quality should be non-decreasing in the number of hearings per proceeding, implying that positive-reacting judges are more attentive. Figure A.28a shows that the two groups of judges again do not differ significantly in the correlation between residualized lower-court grant rates and leniency. Figure A.28b further shows that, for judges who host more hearings per proceeding after surprise reversals, the correlation between residualized reversal risk and leniency is also not significantly different from that of negative-reacting judges. In the unpublished appendix, we show that the results are robust to constructing attentiveness measures at the judge-year level and to alternative classifications based on top-versus-bottom tercile splits.

Taken together, these results suggest that while behavioral diligence, such as hosting more hearings per proceeding or incorporating additional information during hearings, does not substantially alter the reduced form relationship be-

tween leniency and lower-court outcomes, the direction of these responses is consistent with the interpretation that more attentive judges internalize appeal-court feedback and reduce the role of individual bias in their rulings.

5.3.2 Estimating Marginal Treatment Effects

The reduced-form evidence in Section 5.3.1 is consistent with the comparative statics of the model but does not condition on latent case quality. We now take the model directly to the data by estimating marginal treatment effects (MTEs) following the local instrumental variables (LIV) framework of Heckman and Vytlacil (2005) and its empirical implementation in Arnold, Dobbie and Yang (2018). This allows us to measure how reversal probabilities respond to marginal changes in leniency at any point of the leniency/quality distribution, providing a more robust counterpart to the reduced-form patterns.

Let D_{ijt} denote the lower-court grant decision for case i ruled by judge j at time t, and let Z_{ijt} be the constructed leniency instrument. The first-stage relationship defines a propensity score

$$p_{ijt} \equiv \Pr(D_{ijt} = 1 \mid Z_{ijt}, \mathbf{X}_{ijt}) = E[D_{ijt} \mid Z_{ijt}, \mathbf{X}_{ijt}], \tag{5.3}$$

which summarizes how likely a case is to be granted conditional on judge assignment. Section 5.2.2 provides evidence validating the instrument and the strength of this first stage. The MTE is defined as

$$\text{MTE}(p) = \frac{\partial E[\tilde{Y}_{ijt} \mid p_{ijt} = p]}{\partial p}, \tag{5.4}$$

where $\tilde{Y}_{ijt}$ is our outcome of interest—the indicator for whether a denied case is granted on appeal—residualized with the full set of court-by-time fixed effects. Following Heckman, Urzua and Vytlacil (2006), we residualize the lower-court decisions D_{ijt} and leniency Z_{ijt} with the same set of fixed effects. Intuitively, $\text{MTE}(p)$ measures how the expected reversal outcome changes with a marginal increase in the latent grant propensity p, among cases of a given underlying quality p. Since p is constructed as a positive monotone transformation of leniency, the sign and shape of the MTE align with the model's comparative statics in Section 5.1.

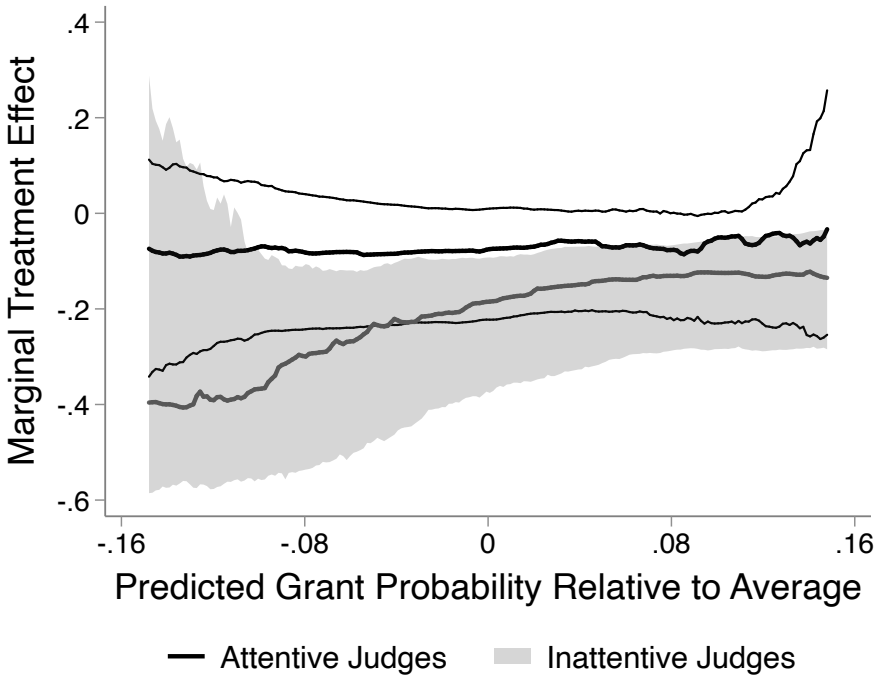
We estimate the MTE in three steps. First, we construct a common grid by

regressing residualized lower-court decisions on judge leniency and computing the predicted value $\tilde{p}_{ijt} = \bar{D}_{ijt} + \hat{\theta} Z_{ijt}$. We then trim $\tilde{p}_{ijt}$ to the 10th–90th percentiles and create a fixed grid of 200 evaluation points that is common to all samples and bootstrap replications. Second, for attentive and inattentive judges separately, we residualize the reversal indicator on the same fixed effects and estimate $E[\tilde{Y}_{ijt} \mid p_{ijt} = p]$ using local linear regression with a bandwidth of 0.08. The derivative with respect to p is obtained numerically. Third, we compute bootstrap confidence intervals by re-sampling at the judge level (500 replications) and constructing reverse-percentile intervals at each grid point. Since greater lower-court leniency corresponds to higher p_{ijt}, MTE(p) can be interpreted as the marginal change in the reversal probability caused by a small increase in leniency, evaluated among cases of latent quality p.

Figure 11 plots the estimated MTEs separately for attentive and inattentive judges, evaluated on a common grid of predicted grant probabilities. For attentive judges, the MTE curve is consistently close to zero across the support of p. This indicates that, conditional on case quality, marginal changes in leniency generate little to no change in reversal rates. This pattern is exactly what the model predicts for judges whose thresholds have converged to the appeal court standard c^* operated in the low sensitivity region. In contrast, inattentive judges exhibit a markedly negative MTE, especially at low values of p associated with stricter effective thresholds. In this strict region, small increases in leniency sharply reduce the probability of reversal, and the MTE is uniformly more negative than for attentive judges. As p increases, the MTE for inattentive judges gradually flattens, consistent with the diminishing mass of borderline strong denials at higher levels of leniency. These results mirror the reduced-form patterns in Figures 9: attentive judges' reversal leniency relationships are nearly flat, while inattentive judges' are steep, consistent with the predictions of Lemma 5.1.

In Figures A.29 and A.30, we plot the distribution of MTEs for subsamples defined by judges' hearing-diligence attentiveness. A similar but weaker pattern emerges: judges who reduce early predictability or host more hearings per proceeding after surprise reversals tend to exhibit MTEs that are close to zero, whereas their less attentive counterparts display more negative MTEs. However, as shown in Figures A.27 and A.28, the differences between the two groups are small in magnitude and generally not statistically significant. Overall, these results suggest that

Figure 11: Marginal Treatment Effect, Grant Rate Attentiveness



Notes: The figure plots the marginal treatment effects (MTEs) of lower-court decisions on lower-court reversals separately for cases ruled by attentive and inattentive judges, classified by grant-rate attentiveness as described in Section 5.2.1. The MTEs are estimated following the procedure outlined in Section 5.3. The black solid line shows the MTEs estimated on the subsample of cases ruled by attentive judges, with the dashed lines as 90 percent confidence intervals. The gray solid line shows those estimated on the subsample of cases ruled by inattentive judges, with the shaded area as 90 percent confidence intervals. Standard errors are computed using 500 bootstrap replications clustered at the judge level.

attentiveness in grant rate behavior is the primary channel through which judges attenuate the influence of personal leniency on reversal outcomes; hearing diligence adjustments move in the same direction but play a more limited role.

We next estimate MTEs separately for the two distinct margins in the reversal process: whether a denied case proceeds to the appeal court at all, and whether an appealed case is granted. Specifically, we decompose Y_{ijt} as

$$Y_{ijt} \equiv A_{ijt} \cdot S_{ijt},$$

where $A_{ijt} = 1$ if the denied case i is appealed (and 0 otherwise) and $S_{ijt} = 1$ if case i's appeal is granted, conditional on $A_{ijt} = 1$. Differentiating with respect to p

yields

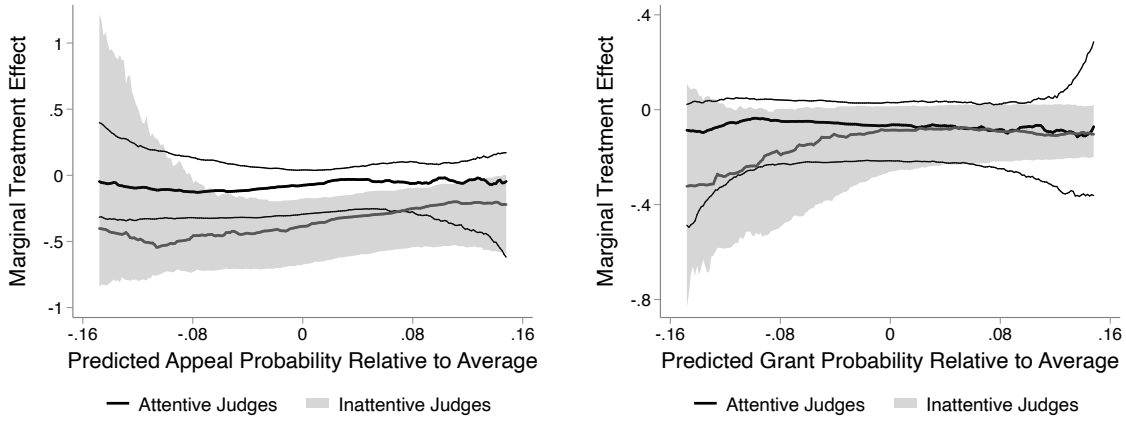
$$\begin{aligned}\text{MTE}_Y(p) &= \frac{\partial}{\partial p} E\left[\tilde{Y} \mid p\right] \\ &= \underbrace{\underbrace{\text{MTE}_A(p)}_{\equiv \frac{\partial}{\partial p} E\left[\tilde{A}|p\right]} \cdot E\left[S \mid A{=}1, p\right]}_{\text{selection margin}} + \underbrace{E\left[A \mid p\right] \cdot \underbrace{\text{MTE}_S(p)}_{\equiv \frac{\partial}{\partial p} E\left[\tilde{S}|A=1,p\right]}}_{\text{success margin}} .\end{aligned} \tag{5.5}$$

We begin with the selection margin. Figure 12a plots the estimated MTEs for the probability that a denied case proceeds to appeal, separately for attentive and inattentive judges. For inattentive judges, the MTE is consistently and significantly negative, especially in the strict-judge region (low p), indicating that greater leniency sharply reduces the likelihood that a denial is appealed. In contrast, the MTE for attentive judges is close to zero across the distribution of p, suggesting that they have learned from past reversals and stabilized their granting threshold to align with the appeal court standard.

We now turn to the intensive margin, which captures variation in appeal success among appealed cases. Figure 12b plots the estimated MTEs for the probability that an appeal is granted, conditional on appeal, for attentive and inattentive judges. For inattentive judges, the MTE is again negative, particularly at low values of p, indicating that stricter judges deny some strong cases that later win on appeal. As judge leniency increases, these strong denials disappear from the appeal pool, lowering the overall success rate of appeals and generating a negative MTE in the lower part of the distribution. The much flatter curve for attentive judges indicates that, conditional on appeal, the quality of their denials is stable across the leniency distribution. This again suggests that attentive judges' denials are close to the appeal court's threshold, leaving little room for systematic correction in the appeal court.

Taken together, the overall MTEs in Figure 11 further confirms the compositional bias operating through both extensive and intensive margins as predicted by our model. Inattentive judges exhibit greater threshold noise: small changes in leniency move many borderline cases across the grant/deny boundary, producing steep negative MTEs. On the contrary, attentive judges learn from surprise reversals, this learning process eliminates most of the compositional bias on both margins, yielding flat and near-zero MTEs.

Figure 12: Marginal Treatment Effect Decomposition on Two Margins



(a) Selection Margin (b) Success Margin

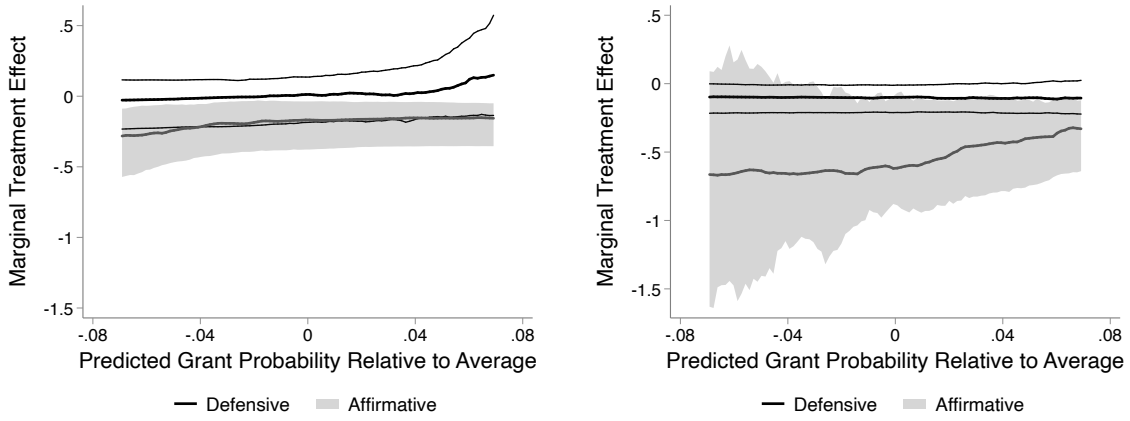
Notes: The figure plots the marginal treatment effects (MTEs) of lower-court decisions on appeals and appeal success separately for cases ruled by attentive and inattentive judges, classified by grant-rate attentiveness as described in Section 5.2.1. The MTEs are estimated following the procedure outlined in Section 5.3. The left panel plots the MTEs on appeal filing among all denied cases, the right panel plots the MTEs on appeal grant, conditional on appealing. In both panels, the black solid line shows the MTEs estimated on the subsample of cases ruled by attentive judges, with the dashed lines as 90 percent confidence intervals. The gray solid line shows those estimated on the subsample of cases ruled by inattentive judges, with the shaded area as 90 percent confidence intervals. Standard errors are computed using 500 bootstrap replications clustered at the judge level.

In Figures A.31 and A.32, we replicate the MTE decomposition for judge groupings based on early predictability attentiveness and hearings per proceeding attentiveness. Consistent with the total MTE patterns shown in Figures A.29 and A.30, differences between attentive and inattentive judges are less pronounced along both the extensive and intensive margins. These findings reinforce that the learning and calibration documented in Figures 11 and 12 do not come at the expense of hearing diligence; rather, more attentive procedural behavior complements the stabilization of decision thresholds aligned with appeal court standards.

Finally, we examine whether the learning patterns documented above vary across the two major procedural categories of asylum adjudication: *defensive* and *affirmative* cases. Defensive cases are filed when the asylum seeker is already in the removal proceeding, typically involving higher-stakes consequences for applicants, more developed evidentiary records and more structured hearings. Affirmative cases are filed voluntarily by asylum seekers that are not in the removal proceeding yet. This procedural difference does not necessarily reflect systematic distinctions in case quality, but it does shape the informational environment sur-

rounding adjudication: When a defensive case is denied, appeals tend to be filed more consistently by applicants with strong claims, generating clearer directional feedback for judges. Affirmative cases, by contrast, exhibit greater heterogeneity in representation and appeal incentives, producing weaker or noisier signals when reversals occur. Under the learning framework in Section 5.1, these differences imply that inattentive judges may update their thresholds more effectively in defensive cases than in affirmative ones, while attentive judges should realign thresholds in both settings.

Figure 13: Marginal Treatment Effects, Affirmative versus Defensive



(a) Attentive Judges (b) Inattentive Judges

Notes: The figure plots total marginal treatment effects (MTEs) for defensive and affirmative proceedings, separately for attentive and inattentive judges. Both panels use the common grid and estimation procedure described in Section 5.3.2. In both panels, the black solid line shows the MTEs estimated on the subsample of cases ruled by attentive judges, with the dashed lines as 90 percent confidence intervals. The gray solid line shows those estimated on the subsample of cases ruled by inattentive judges, with the shaded area as 90 percent confidence intervals. Standard errors are computed using 500 bootstrap replications clustered at the judge level.

Figure 13 plots the total MTEs for affirmative and defensive proceedings separately for attentive and inattentive judges. Among attentive judges, MTEs are uniformly close to zero for both case types, indicating that these judges have moved their effective thresholds toward the appeal court standard c^* across procedural contexts. Among inattentive judges, defensive cases also display relatively flat MTEs, consistent with the idea that the clearer feedback generated by defensive denials helps even inattentive judges converge to c^*. In contrast, inattentive judges exhibit sharply negative MTEs in the strict region for affirmative cases, which indicates the inattentive judges' lack of learning or affirmative cases.

Taken together, these patterns reinforce the learning mechanism of the model. Judges appear to allocate more attention to defensive cases, and, as a result, even inattentive judges tend to update their defensive thresholds. For affirmative cases, inattentive judges fail to update while attentive judges converge to the appeal court standard.

In summary, our marginal treatment effect estimates provide a structural counterpart to the reduced form results in Section 5.3.1. Across all specifications, attentive judges exhibit MTEs that are flat and close to zero, indicating that once their thresholds move toward the appeal court standard c^*, marginal changes in leniency have virtually no effect on reversals. Inattentive judges, by contrast, display steep and negative MTEs in the strict region, reflecting persistent threshold misalignment and a large mass of borderline strong cases whose outcomes remain highly sensitive to marginal shifts in leniency. The extensive intensive decomposition shows that these patterns operate through both appeal filing and appeal success. The case type analysis further demonstrates that this learning mechanism depends critically on attention: defensive cases generate strong signals that induce convergence even among inattentive judges. Overall, the evidence aligns closely with the model's predictions: learning from appeal court feedback occurs when judges attend to evaluation signals, and attentiveness is what drives convergence toward consistent decision standards.

6 Conclusion

In this paper, we study how immigration judges learn from appeal court feedback in an environment characterized by substantial discretion, heterogeneous decision styles, and noisy review.

Using comprehensive EOIR and BIA administrative records, we construct high dimensional prediction models that recover judges' expected reversal probabilities and allow us to isolate surprise reversals which deviate sharply from lower court judges' anticipation. These surprise reversals serve as plausibly exogenous information shocks that induce significant changes in judicial decision making: Our event study results show that, following a surprise reversal, grant rates increase by 1.57 percentage points in the first week, and the adjustment persists over subsequent weeks. Importantly, we find no evidence of reduced diligence.

To interpret these findings, we develop a simple model of attention based belief updating. The model predicts a clear division between attentive judges, who place greater weight on surprising signals and move decisively toward the appeal court standard, and inattentive judges, whose decision thresholds remain misaligned. Both reduced-form evidence and marginal treatment effect (MTE) estimation strongly support these predictions: Attentive judges exhibit near-zero marginal reversal elasticity, while inattentive judges' reserval ratee remain sensitive to small changes in leniency. Learning operates through both the extensive margin of appeal filing and the intensive margin of appeal success. We further highlight procedural heterogeneity: defensive cases, which tend to attract more judicial attention, induce convergence even among inattentive judges, while affirmative cases generate meaningful updating only among attentive judges.

Together, these findings point to an attention-based mechanism through which appeal courts discipline lower courts. This perspective reconciles the coexistence of persistent cross-judge disparities with observed post-reversal adjustments, and it underscores the importance of understanding the cognitive foundations of professional learning in low-feedback environments. Our results suggest that raising judges' attention to informative feedback is crucial for the judicial hierarchy to function efficiently and reduce administrative burdens of avoidable appeals.

For future work, one promising direction is to study double learning: how judges integrate experiential learning from their own caseloads and evaluative learning from appeal court decisions. A second direction concerns collective learning: whether judges working in the same court cluster diffuse information or imitate one another's responses to feedback. More broadly, the empirical strategy we have developing, using prediction-based surprise measures to detect learning, offers a general approach for studying other expert decision makers whose decisions are shaped by incomplete and irregular feedback.

In summary, this paper demonstrates that immigration judges do learn from reversals, but only when the information provided is sufficiently unexpected to command attention. The institutional design of review systems should therefore account not only for accuracy and incentives, but also for the cognitive channels through which decision-makers process information. Even modest improvements in the salience and attention could yield meaningful gains in consistency and fairness across a wide range of expert decision environments.

References

Abrams, David S, Marianne Bertrand, and Sendhil Mullainathan. 2012. "Do judges vary in their treatment of race?" *The Journal of Legal Studies*, 41(2): 347–383.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association*, 91(434): 444–455.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial bias in bail decisions." *The Quarterly Journal of Economics*, 133(4): 1885–1932.

Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro HC Sant'Anna. 2025. "Difference-in-differences designs: A practitioner's guide." *arXiv preprint arXiv:2503.13323*.

Bastianello, Francesca, and Alex Imas. 2025. "Biases in belief updating within and across domains." *Available at SSRN 5413920*.

Bhuller, Manudeep, and Henrik Sigstad. 2025. "Feedback and learning: The causal effects of reversals on judicial decision-making." *Review of Economic Studies*, 92(4): 2359–2397.

Bindler, Anna, and Randi Hjalmarsson. 2019. "Path dependency in jury decision making." *Journal of the European Economic Association*, 17(6): 1971–2017.

Boyd, Christina L. 2015. "The hierarchical influence of courts of appeals on district courts." *The Journal of Legal Studies*, 44(1): 113–141.

Breiman, Leo. 2001. "Random forests." *Machine learning*, 45(1): 5–32.

Buser, Thomas, Leonie Gerhards, and Joël Van Der Weele. 2018. "Responsiveness to feedback as a personal trait." *Journal of Risk and Uncertainty*, 56(2): 165–192.

Cameron, Charles M, Jeffrey A Segal, and Donald Songer. 2000. "Strategic auditing in a political hierarchy: An informational model of the Supreme Court's certiorari decisions." *American Political Science Review*, 94(1): 101–116.

Capozza, Francesco, Ingar Haaland, Christopher Roth, and Johannes Wohlfart. 2021. "Studying information acquisition in the field: A practical guide and review." CEBI Working Paper Series.

Chen, Daniel L. 2023. "Judicial compliance in district courts." *International Review of Law and Economics*, 74: 106122.

Chen, Daniel L, and Arnaud Philippe. 2023. "Clash of norms judicial leniency on defendant birthdays." *Journal of Economic Behavior & Organization*, 211: 324–344.

Chen, Daniel L, Tobias J Moskowitz, and Kelly Shue. 2016. "Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires." *The Quarterly Journal of Economics*, 131(3): 1181–1242.

Dobbie, Will, Jacob Goldin, and Crystal S Yang. 2018. "The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges." *American Economic Review*, 108(2): 201–240.

Dunn, Matt, Levent Sagun, Hale Şirin, and Daniel Chen. 2017. "Early predictability of asylum court decisions." 233–236.

Enke, Benjamin, and Thomas Graeber. 2023. "Cognitive uncertainty." *The Quarterly Journal of Economics*, 138(4): 2021–2067.

Eren, Ozkan, and Naci Mocan. 2018. "Emotional judges and unlucky juveniles." *American Economic Journal: Applied Economics*, 10(3): 171–205.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. "Do we need hundreds of classifiers to solve real world classification problems?" *The journal of machine learning research*, 15(1): 3133–3181.

Gennaioli, Nicola, and Andrei Shleifer. 2008. "Judicial fact discretion." *The Journal of Legal Studies*, 37(1): 1–35.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2025. "Leniency Designs: An Operator's Manual." National Bureau of Economic Research.

Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. 2022. "Why do tree-based models still outperform deep learning on typical tabular data?" *Advances in neural information processing systems*, 35: 507–520.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2014. "Learning through noticing: Theory and evidence from a field experiment." *The Quarterly Journal of Economics*, 129(3): 1311–1353.

Heckman, James J, and Edward Vytlacil. 2005. "Structural equations, treatment effects, and econometric policy evaluation 1." *Econometrica*, 73(3): 669–738.

Heckman, James J, Sergio Urzua, and Edward Vytlacil. 2006. "Understanding instrumental variables in models with essential heterogeneity." *The review of economics and statistics*, 88(3): 389–432.

Heyes, Anthony, and Soodeh Saberian. 2019. "Temperature and decisions: evidence from 207,000 court cases." *American Economic Journal: Applied Economics*, 11(2): 238–265.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human decisions and machine predictions." *The quarterly journal of economics*, 133(1): 237–293.

Klein, David E, and Robert J Hume. 2003. "Fear of reversal as an explanation of lower court compliance." *Law & Society Review*, 37(3): 579–581.

Philippe, Arnaud, and Aurélie Ouss. 2018. ""No hatred or malice, fear or affection": Media and sentencing." *Journal of Political Economy*, 126(5): 2134–2178.

Rachlinski, Jeffrey J, and Andrew J Wistrich. 2017. "Judging the judiciary by the numbers: Empirical research on judges." *Annual Review of Law and Social Science*, 13(1): 203–229.

Ramji-Nogales, Jaya, Andrew I Schoenholtz, and Philip G Schrag. 2007. "Refugee roulette: Disparities in asylum adjudication." *Stan. L. Rev.*, 60: 295.

Schwartzstein, Joshua. 2014. "Selective attention and learning." *Journal of the European Economic Association*, 12(6): 1423–1452.

Shayo, Moses, and Asaf Zussman. 2011. "Judicial ingroup bias in the shadow of terrorism." *The Quarterly journal of economics*, 126(3): 1447–1484.

Spitzer, Matthew, and Eric Talley. 2013. "Left, right, and center: Strategic information acquisition and diversity in judicial panels." *The Journal of Law, Economics, & Organization*, 29(3): 638–680.

Westerland, Chad, Jeffrey A Segal, Lee Epstein, Charles M Cameron, and Scott Comparato. 2010. "Strategic defiance and compliance in the US courts of appeals." *American Journal of Political Science*, 54(4): 891–905.

A Figures and Tables

Table A.1: Top Countries of Origin in Asylum Proceedings, 1986–2019

Country of Origin	*Number of Proceedings*
China	140,783
El Salvador	88,869
Guatemala	70,506
Mexico	51,210
Haiti	48,022
Honduras	44,432
Colombia	29,420
India	28,597
Nicaragua	22,173
Albania	13,104

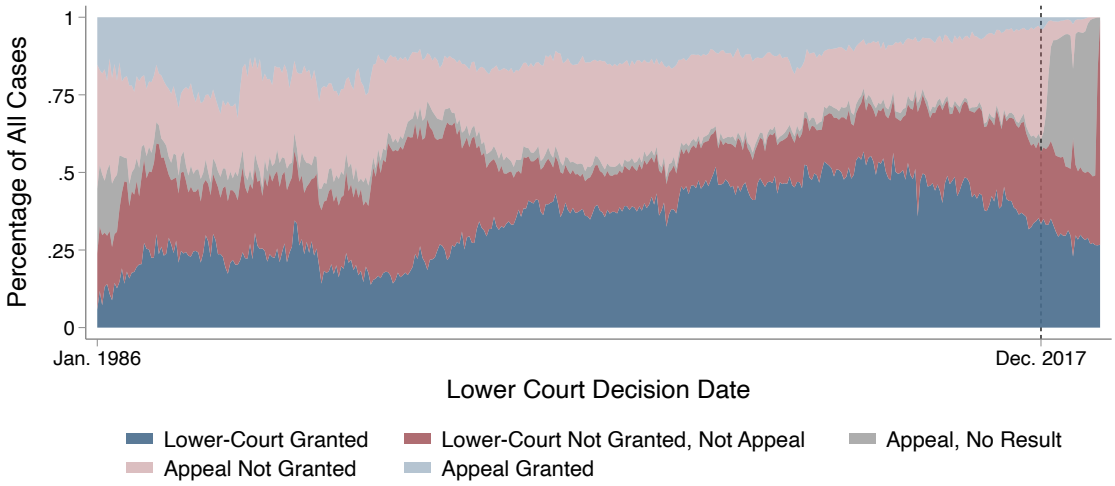
Note: Counts reflect unique court proceedings in EOIR administrative data, 1986–2019.

Table A.2: Top U.S. States by Number of Asylum Hearings, 1986–2019

Hearing State	*Number of Proceedings*
New York (NY)	201,353
California (CA)	173,029
Florida (FL)	121,238
Texas (TX)	59,301
Maryland (MD)	26,212
New Jersey (NJ)	25,744
Virginia (VA)	25,213
Massachusetts (MA)	18,754
Illinois (IL)	18,003
Pennsylvania (PA)	16,696

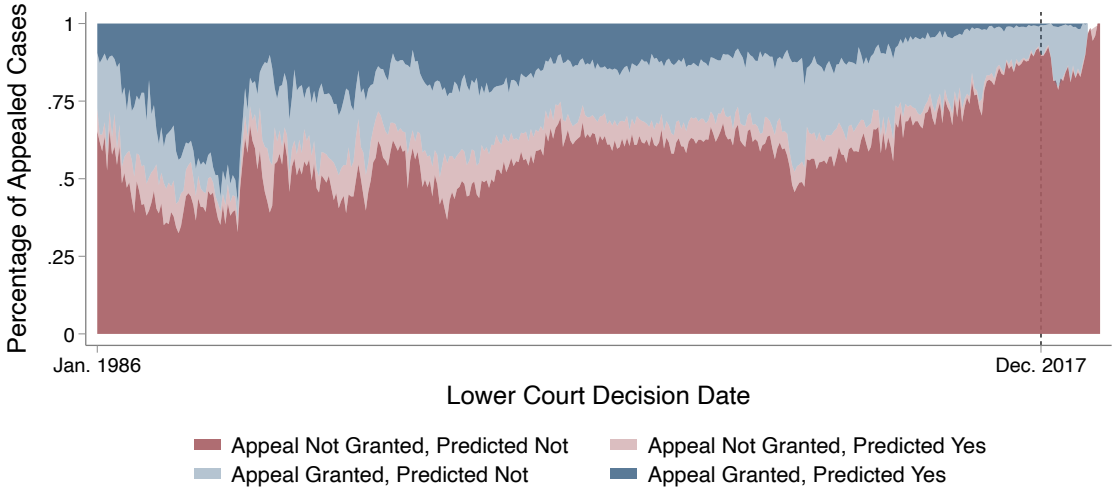
Note: Counts reflect unique court proceedings in EOIR administrative data, 1986–2019. State abbreviations follow EOIR hearing location codes.

Figure A.1: Share of Asylum Cases by Adjudication Outcome, 1986–2019



Notes: The figure displays the distribution of asylum case outcomes over the sample period, based on all proceedings in the EOIR administrative data between 1986 and 2019. Each area represents the share of cases that fall into one of five categories: (i) granted by the lower court; (ii) denied by the lower court without appeal; (iii) denied and appealed but no result; (iv) appealed and denied by the BIA; and (v) appealed and granted by the BIA. Percentages are computed relative to the total number of cases decided in the lower court in a given month.

Figure A.2: Share of Appeals by Realized and Predicted Outcome, 1986–2019



Notes: The figure displays the distribution of realized and predicted appeal outcomes over the sample period, based on all appeals in the BIA data from 1986 to 2019. Each area represents the share of cases falling into one of four categories: (i) appeal not granted and predicted not granted (true negatives); (ii) appeal not granted but predicted granted (false positives); (iii) appeal granted but predicted not granted (false negatives); and (iv) appeal granted and predicted granted (true positives). Percentages are computed relative to the total number of cases decided in the lower court in a given month.

Table A.3: BIA Decision Classes

Category	Decision Type
Positive	Background Check Remand
	Grant With No Remand
	Granted
	Remand
	Sustain
	Temporary Protected Status
	Termination
Negative	Denied
	Dismiss Appeal/Affirm IJ's Decision
	Dismissed (Grant V/D 30 days)
	Dismissed (Voluntary Departure Granted)
	Rejection
	Summary Affirmance
	Summary Dismiss
	Summary Dismissal (O) Other
	Summary Dismissal (a) inad reason on appeal
	Summary Dismissal - Both (a) & (e)
Neutral	Summary Dismissal (e) no brief
	Administrative Return
	Dismiss as Untimely
	Lacks Jurisdiction - BIA
	Moot Bond
	Withdrawal of Appeal
	Continued
	Other
	Other Soriano

Note: This table lists 27 distinct BIA decision outcomes grouped into three categories used in the analysis: positive (granted), negative (dismissed), and neutral. Classification follows EOIR BIA decision codes.

Table A.4: Summary of Judge Characteristics

Variable	**Mean**
Male	0.617
Years between law school graduation and appointment	18.821
Years between college and law school graduation	5.195
Private-practice experience (yes/no)	0.437
Military experience (yes/no)	0.173
NGO experience (yes/no)	0.129
INS experience (yes/no)	0.342
Other government experience (yes/no)	0.686
Academic experience (yes/no)	0.069

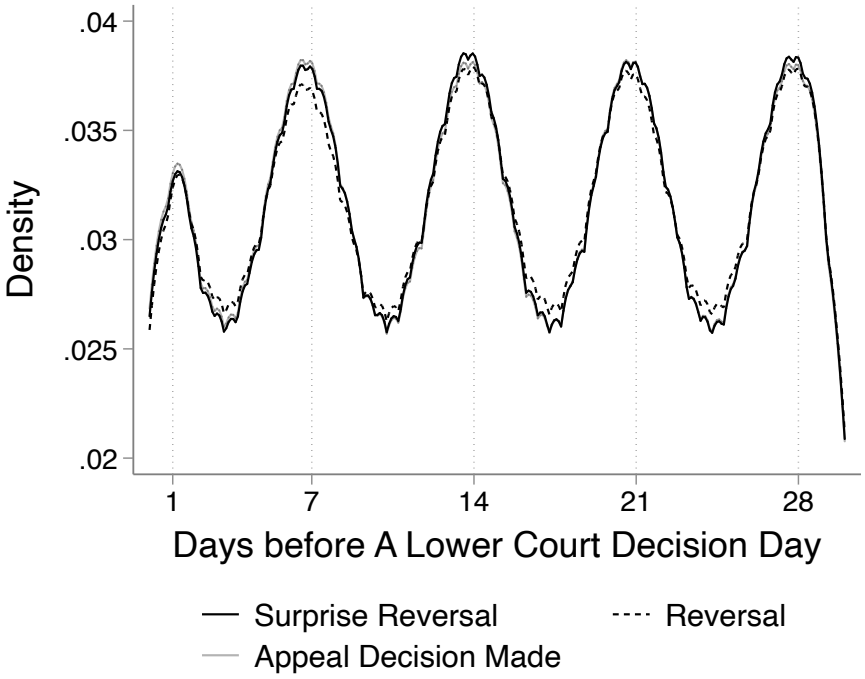
Note: This table reports the mean of selected judge characteristics based on 698 EOIR judges.

Table A.5: Predictive Performance of Early-Predictability Models

Model	**ROC AUC**	**Accuracy**	**Log Loss**
Judge ID only	0.7406	0.7437	0.5884
Judge ID + Nationality	0.7967	0.7699	0.7142
Judge ID + Opening Date	0.6915	0.7039	0.7918
Judge ID + Nationality + Opening YM	0.7918	0.7625	0.5541
Judge ID + Nationality + Opening YM + Filing YM	0.7963	0.7669	0.5038

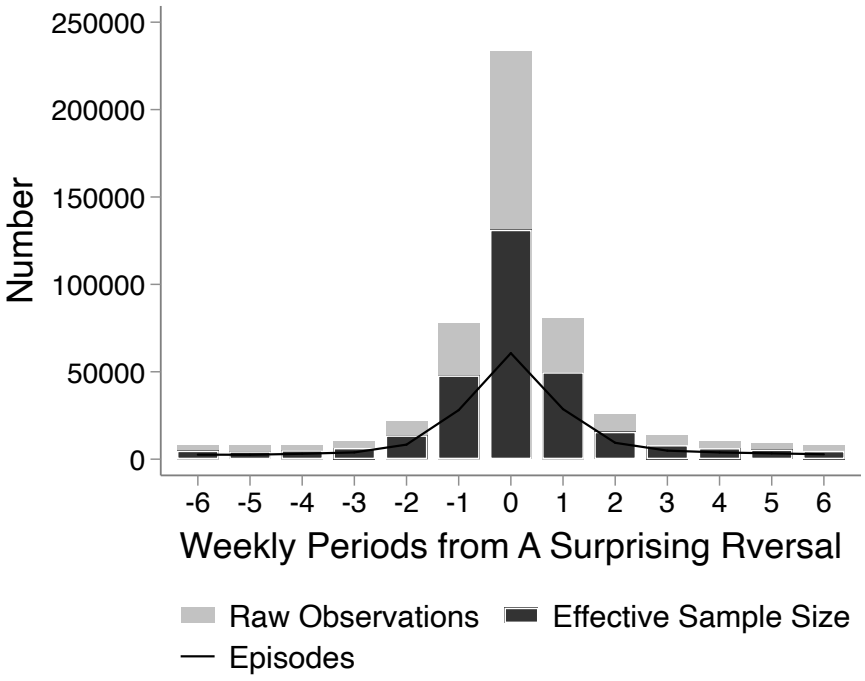
Notes: Each row reports the predictive performance of one of the five early-predictability models described in Section 3.1. Metrics include the area under the receiver operating characteristic curve (ROC AUC), classification accuracy, and log loss. All models are trained sequentially using random-forest estimators with rolling-year cross-validation.

Figure A.3: Arrival Probability of Appeal Decisions



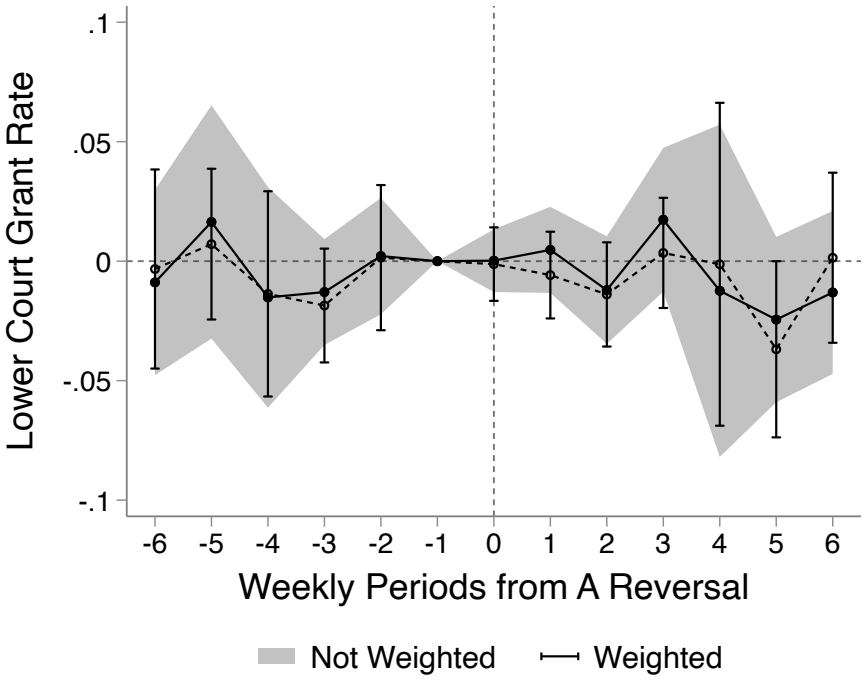
Notes: The figure plots kernel density estimates of the timing of appeal decisions in the 30 days preceding a lower-court ruling. Reversals and surprise reversals are defined as described in Section 4. Densities are estimated using an Epanechnikov kernel with a bandwidth of 1, based on all lower-court observations in the sample.

Figure A.4: Nearest Event Attribution Trimmed Sample Distribution



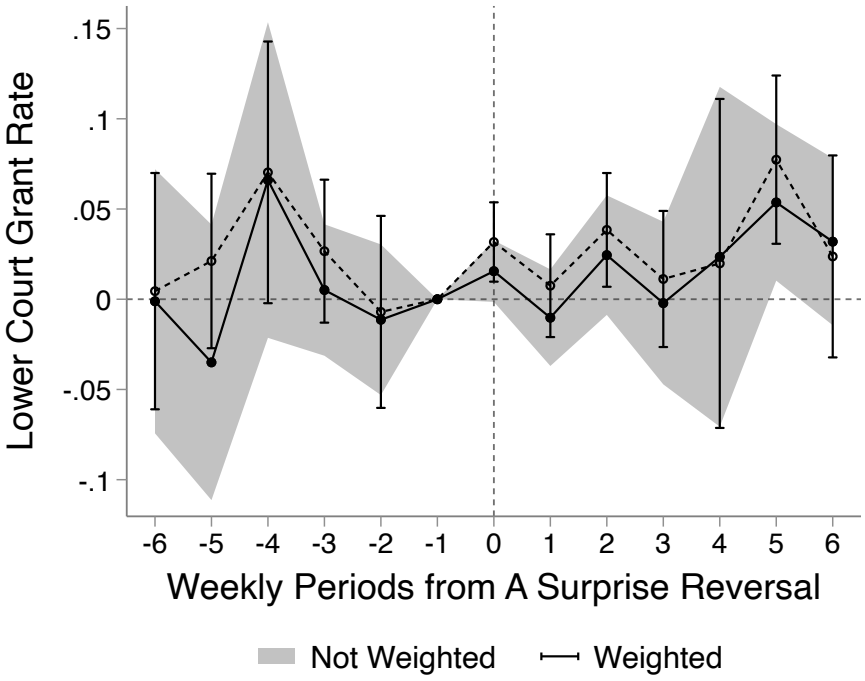
Notes: The figure plots the distribution of observations across event periods of the nearest event attribution trimmed sample described in Section 4. Raw observations refer to the raw count of case-event-period observation in each period, effective sample size is calculated as desribed in Section 4. Episodes refer to the number of reversals that have observations of certain periods.

Figure A.5: Effect of Reversals on Lower Court Grant Rate, Unbinned



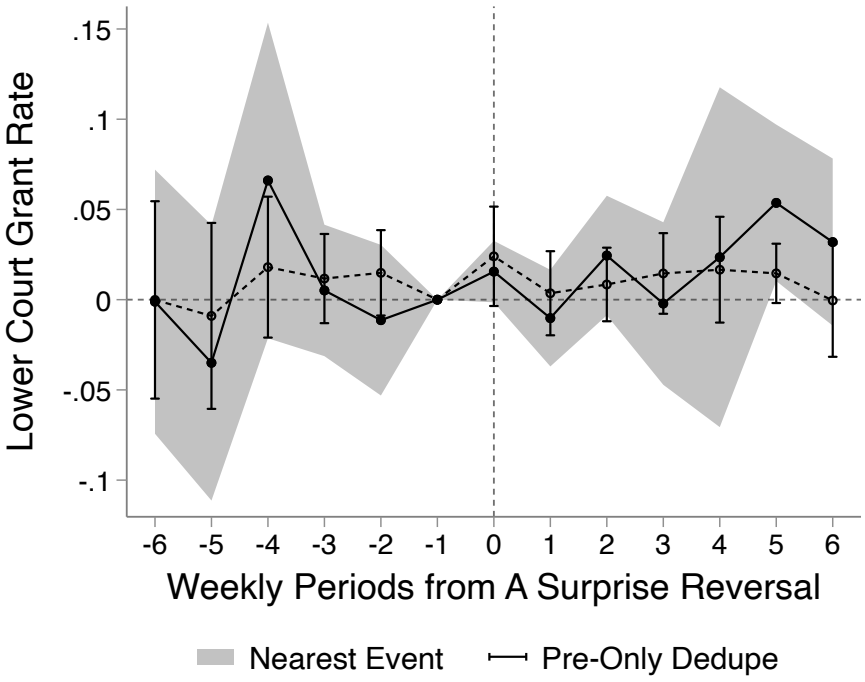
Notes: The figure plots plots $\hat{\beta}_k$, estimated coefficients of the interaction term of the period indicators and the reversal indicator of Specification (4.1) with unbinned periods indexed by $K = [-6, 6]$, comparing unweighted and event-equal weighted estimates. The sample includes all appeal decisions from 1986 to 2019 as events, trimmed using the nearest-event attribution method described in Section 4. Weights are calculated on the estimation sample as described in Section 4.1. The x-axis indexes seven-day periods around each appeal decision. All estimations include controls for appeal fixed effects and appeal ruling calendar time fixed effects. Standard errors are clustered at the judge level. Unweighted estimates are shown as solid dots with shaded 95% confidence intervals, weighted estimates are shown as open circles with bracketed 95% confidence intervals.

Figure A.6: Effect of Surprise Reversals on Lower Court Grant Rate, Unbinned



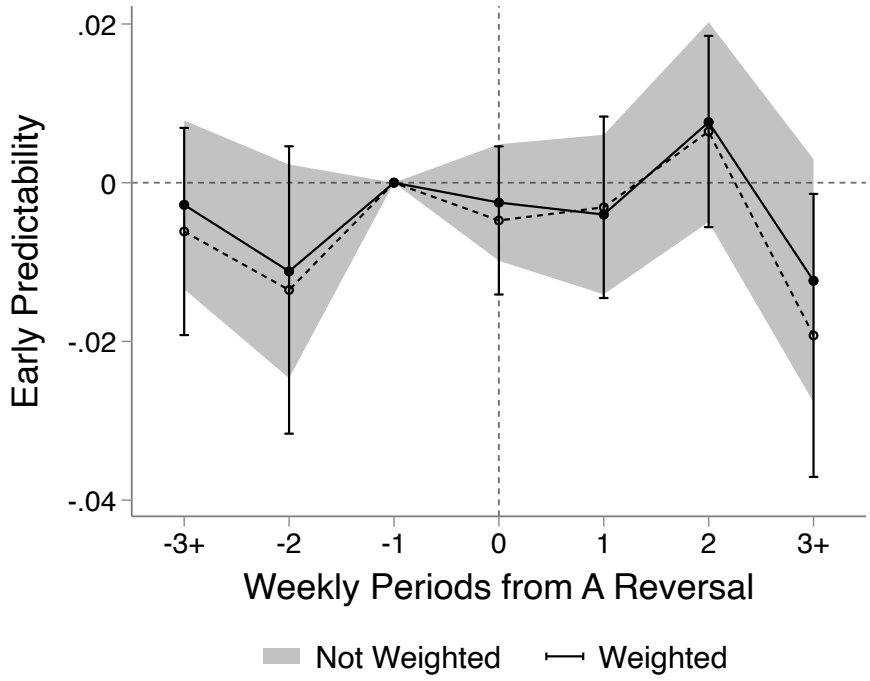
Notes: The figure plots plots $\hat{b}_k$, estimated coefficients of the interaction term of the period indicators and the surprise reversal indicator of Specification (4.2) with unbinned periods indexed by $K = [-6, 6]$, comparing unweighted and event-equal weighted estimates. The sample includes all appeal decisions from 1986 to 2019 as events, trimmed using the nearest-event attribution method described in Section 4. Weights are calculated on the estimation sample as described in Section 4.1. The x-axis indexes seven-day periods around each appeal decision. All estimations include controls for appeal fixed effects and appeal ruling calendar time fixed effects. Standard errors are clustered at the judge level. Unweighted estimates are shown as solid dots with shaded 95% confidence intervals, weighted estimates are shown as open circles with bracketed 95% confidence intervals.

Figure A.7: Effect of Surprise Reversals on Lower Court Grant Rate, Sampling Trimming



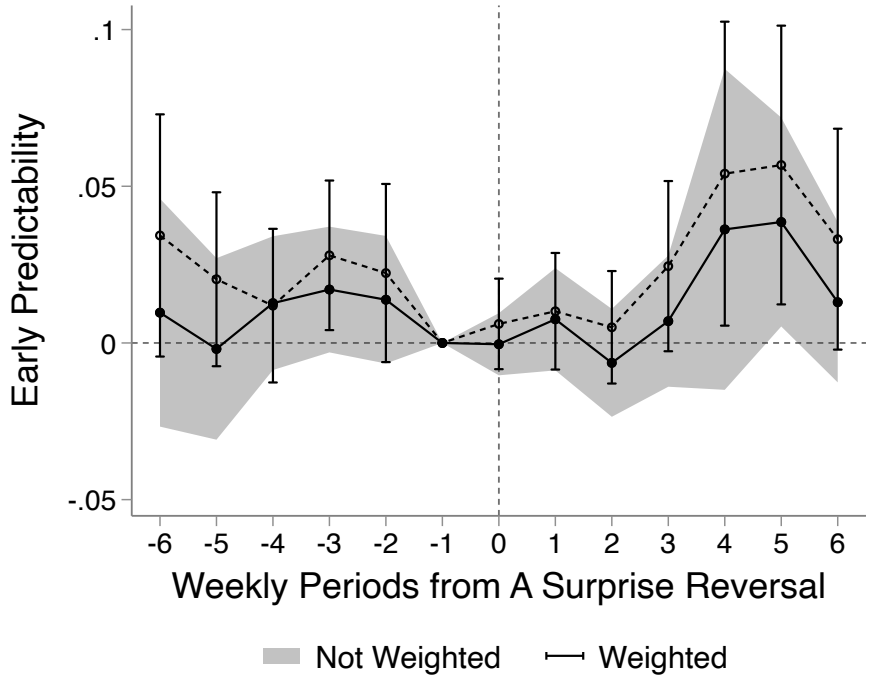
Notes: The figure plots plots $\hat{b}_k$, unweighted estimated coefficients of the interaction term of the period indicators and the surprise reversal indicator of Specification (4.2) with unbinned periods indexed by $K = [-6, 6]$. The sample includes all appeal decisions from 1986 to 2019 as events, trimmed using the nearest-event attribution method and the pre-only dedupe method described in Section 4. All estimations include controls for appeal fixed effects and appeal ruling calendar time fixed effects. Standard errors are clustered at the judge level. Estimates on the nearest-event attribution trimmed sample are shown as solid dots with shaded 95% confidence intervals, estimates on the preonly-dedupe trimmed sample are shown as open circles with bracketed 95% confidence intervals.

Figure A.8: Effect of Reversals on Lower Court Early Predictability



Notes: The figure plots $\hat{\beta}_k$, estimated coefficients of the interaction term of the period indicators and the reversal indicator of Specification (4.1), with the continous early predictability measure described in Section 4.3 as the outcome variable. The sample includes all appeal decisions from 1986 to 2019 as events, trimmed using the nearest-event attribution method described in Section 4.1. Weights are calculated on the estimation sample as described in Section 4.1. The x-axis indexes seven-day periods around each appeal decision. All estimations include controls for appeal fixed effects and appeal ruling calendar time fixed effects. Standard errors are clustered at the judge level. Unweighted estimates are shown as solid dots with shaded 95% confidence intervals, weighted estimates are shown as open circles with bracketed 95% confidence intervals.

Figure A.9: Effect of Surprise Reversals on Lower Court Early Predictability, Unbinned



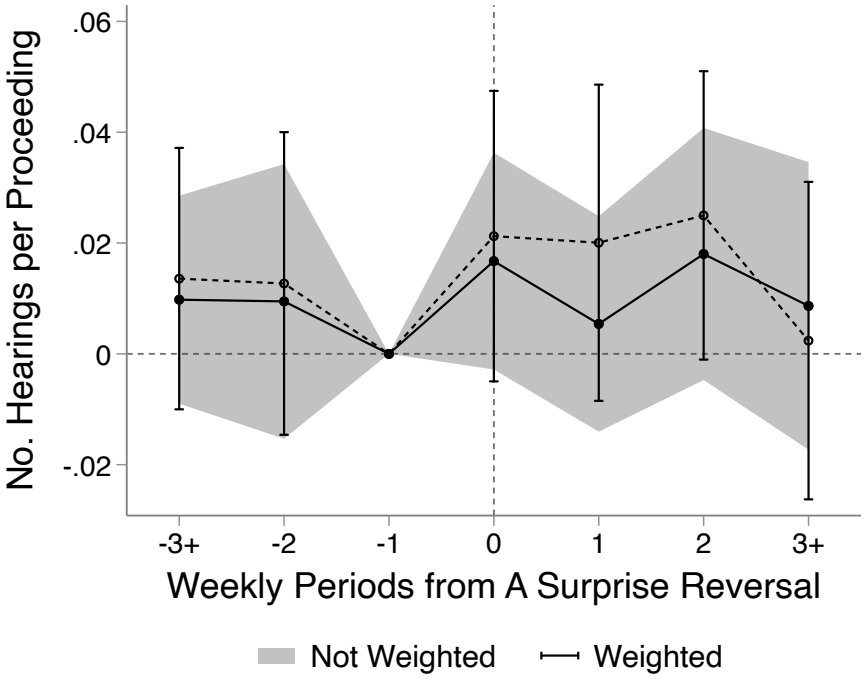
Notes: The figure plots $\hat{b}_k$, estimated coefficients of the interaction term of the period indicators and the surprise reversal indicator of Specification (4.2) with unbinned periods indexed by $K = [-6, 6]$ and the continous early predictability measure described in Section 4.3 as the outcome variable, comparing unweighted and event-equal weighted estimates. The sample includes all appeal decisions from 1986 to 2019 as events, trimmed using the nearest-event attribution method described in Section 4. Weights are calculated on the estimation sample as described in Section 4.1. The x-axis indexes seven-day periods around each appeal decision. All estimations include controls for appeal fixed effects and appeal ruling calendar time fixed effects. Standard errors are clustered at the judge level. Unweighted estimates are shown as solid dots with shaded 95% confidence intervals, weighted estimates are shown as open circles with bracketed 95% confidence intervals.

Table A.6: Hearing Diligence Measures, 1984-2019

Measure	Mean	Median	N
Panel A: Proceedings related to the main sample			
No. hearings per proceeding	3.79	3	1,227,280
Average duration of hearings (min.)	139.18	122	1,079,585
Share of preparation adjournments	0.08	0	1,103,956
Share of individual calendar hearing	0.35	0.33	1,227,053
Panel B: All proceedings			
No. hearings per proceeding	2.75	2	8,604,714
Average duration of hearings	104.76	70	6,433,513
Share of preparation adjournments	0.09	0	6,471,725
Share of individual calendar hearings	0.16	0	8,599,564

Notes: This table reports summary statistics for the four hearing diligence measures described in Section 4.3. Panel A restricts the sample to proceedings associated with cases in the main analysis, while Panel B includes all proceedings between 1984 and 2019. The measures capture, respectively, the number of hearings per proceeding, the average duration of hearings (winsorized at top 1% level before calculation), the share of adjournments granted for preparation, and the share of hearings held on individual calendars.

Figure A.10: Effect of Surprise Reversals on Hearings per Proceeding



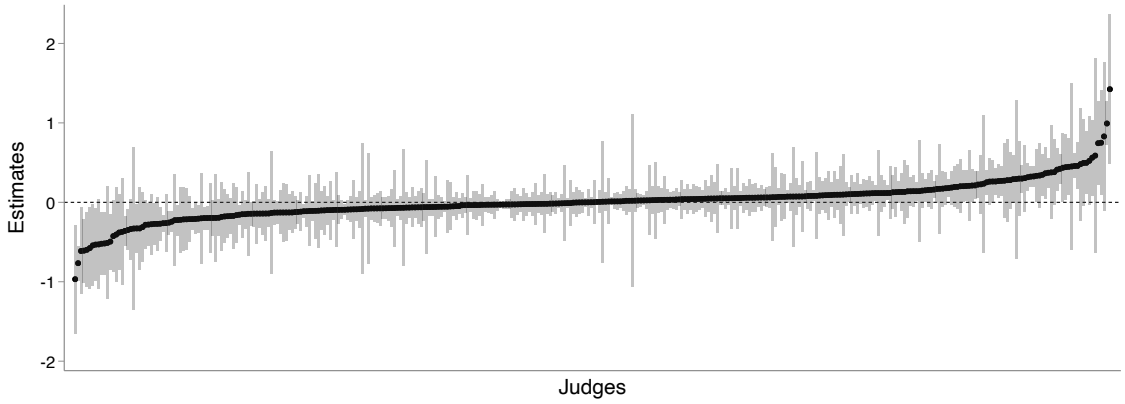
Notes: The figure plots $\hat{\delta}_k$, estimated coefficients of the interaction term of the period indicators and the surprise reversal indicator of Specification (4.3), with the number of hearings per proceeding described in Section 4.3 as the outcome variable, comparing unweighted and event-equal weighted estimates. The sample includes all appeal decisions from 1986 to 2019 as events, trimmed using the nearest-event attribution method described in Section 4. Weights are calculated on the estimation sample as described in Section 4.1. The x-axis indexes seven-day periods around each appeal decision. All estimations include controls for appeal fixed effects and appeal ruling calendar time fixed effects. Standard errors are clustered at the judge level. Unweighted estimates are shown as solid dots with shaded 95% confidence intervals, weighted estimates are shown as open circles with bracketed 95% confidence intervals.

Table A.7: Effect of Surprise Reversals on Lower Court General Hearing Diligence

	(1) No. hearings	(2) Avg. duration	(3) Prep. Share	(4) Ind. Share
$\mathbf{1}(k \in [-6,-3])$	-0.001	-0.963	-0.003	0.001
	(0.002)	(1.599)	(0.005)	(0.008)
$\mathbf{1}(k = -2)$	0.000	-0.351	0.003	-0.001
	(0.002)	(2.090)	(0.008)	(0.010)
$\mathbf{1}(k = 0)$	-0.001	0.133	-0.001	-0.017*
	(0.003)	(1.522)	(0.005)	(0.009)
$\mathbf{1}(k = 1)$	-0.001	0.081	0.001	-0.012
	(0.003)	(1.658)	(0.006)	(0.012)
$\mathbf{1}(k = 2)$	0.001	-1.474	-0.003	-0.009
	(0.002)	(1.883)	(0.008)	(0.010)
$\mathbf{1}(k \in [3,6])$	-0.004	-1.929	0.001	-0.019**
	(0.003)	(1.822)	(0.007)	(0.009)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k \in [-6,-3])$	0.003	1.449	0.003	-0.000
	(0.002)	(1.481)	(0.005)	(0.008)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k = -2)$	0.001	-0.059	-0.003	-0.001
	(0.002)	(1.932)	(0.007)	(0.009)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k = 0)$	-0.000	-0.641	0.002	0.012
	(0.003)	(1.425)	(0.004)	(0.008)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k = 1)$	0.000	0.229	0.001	0.011
	(0.002)	(1.523)	(0.005)	(0.011)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k = 2)$	0.000	1.343	0.004	0.013
	(0.002)	(1.714)	(0.007)	(0.009)
$\mathbf{1}(\text{Surp. Rev.}_e = 1) \times \mathbf{1}(k \in [3,6])$	0.004	0.499	-0.000	0.016*
	(0.003)	(1.688)	(0.007)	(0.008)
adj. R^2	0.217	0.417	0.217	0.171
N	911,110	882,764	852,820	911,043

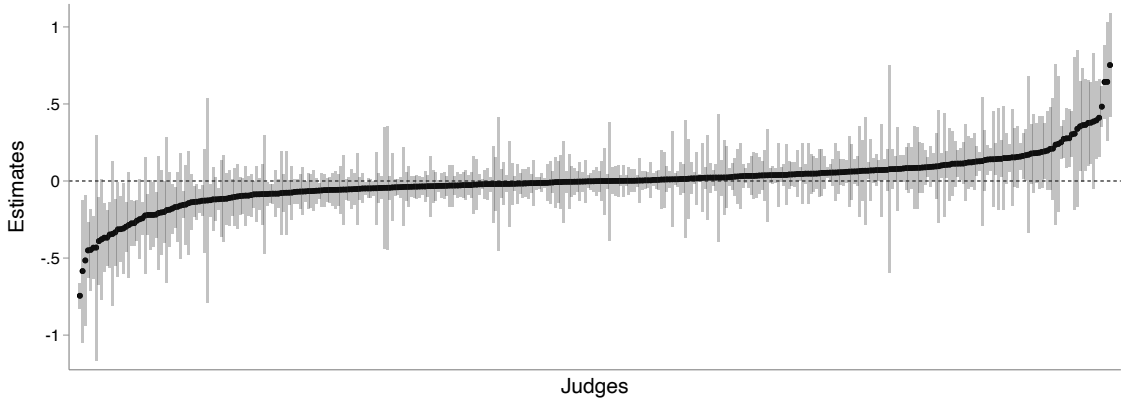
Note: This table reports $\hat{\theta}_k$ and $\hat{\delta}_k$, estimated coefficients of the period indicators and the interaction term of the period indicators and the surprise reversal indicator of Specification (4.2), respectively. Each column presents results using one of the four proceeding-level hearing diligence measures constructed in Section 4.3 on all hearings: the number of hearings, the average duration of adjournments, the share of preparation adjournments, and the share of individual calendar hearings. The sample includes all appeal decisions from 1986 to 2019 as events, trimmed using the nearest-event attribution method described in Section 4.1. All estimations are unweighted and include controls for appeal fixed effects and appeal ruling calendar-time fixed effects. Standard errors clustered at the judge level are reported in parentheses. Asterisks denote significance at the 1% (***), 5% (**), and 10% (*) levels.

Figure A.11: Grant Rate Reactions of Lower Court Judges



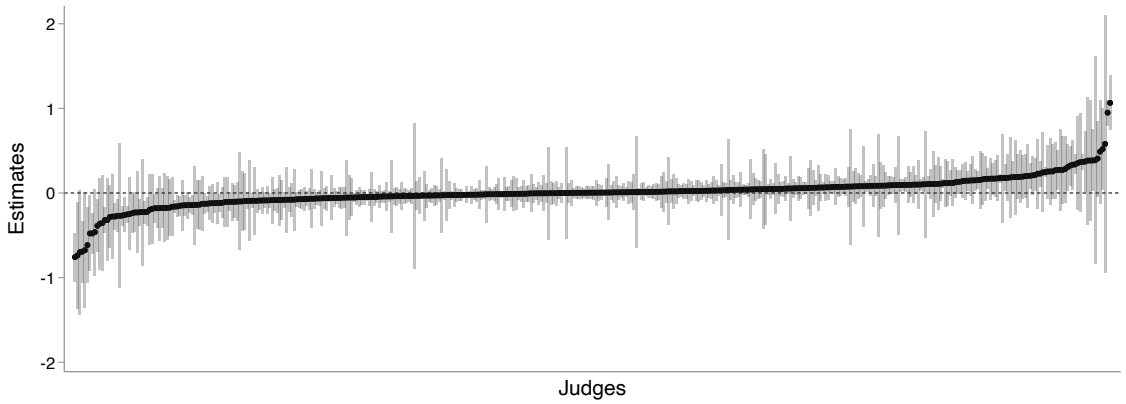
Notes: The figure plots the estimated coefficients $\hat{\mu}_2$ from Specification (5.1), the estimated coefficient of the interaction between the surprise reversal indicator and the period indicator for the first two periods following a surprise reversal. The outcome variable is an indicator for whether a case is granted. The sample includes all appeal decisions from 1986 to 2019, trimmed using the nearest-event attribution method described in Section 4.1. All estimations are unweighted and include appeal fixed effects and appeal ruling calendar-time fixed effects. Standard errors are robust to heteroskedasticity. Each dot represents an estimate for one of 354 judges, ranked by the magnitude of $\hat{\mu}_2$; gray lines denote 95% confidence intervals.

Figure A.12: Early Predictability Reactions of Lower Court Judges



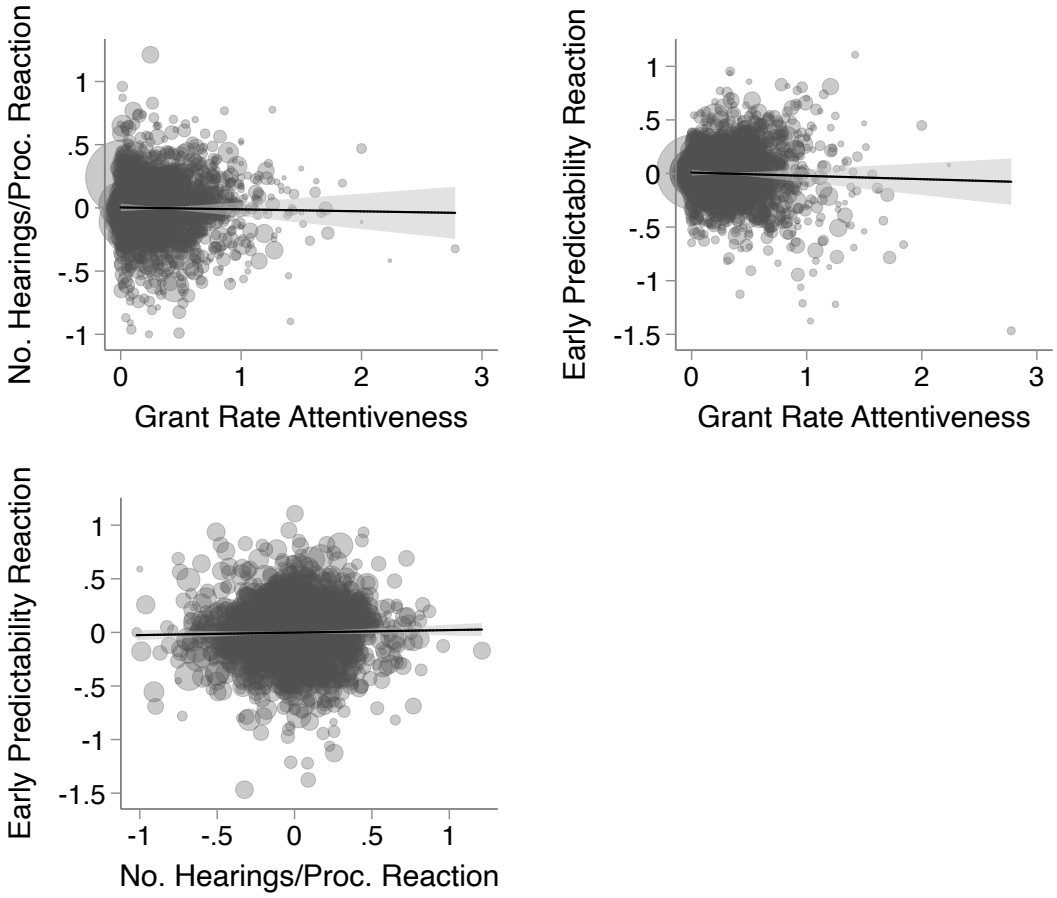
Notes: The figure plots the estimated coefficients $\hat{\mu}_2$ from Specification (5.1), the estimated coefficient of the interaction between the surprise reversal indicator and the period indicator for the first two periods following a surprise reversal. The outcome variable is the continuous early predictability score described in Section 4.3. The sample includes all appeal decisions from 1986 to 2019, trimmed using the nearest-event attribution method described in Section 4.1. All estimations are unweighted and include appeal fixed effects and appeal ruling calendar-time fixed effects. Standard errors are robust to heteroskedasticity. Each dot represents an estimate for one of 416 judges, ranked by the magnitude of $\hat{\mu}_2$; gray lines denote 95% confidence intervals.

Figure A.13: No. Hearings per Proceeding Reactions of Lower Court Judges



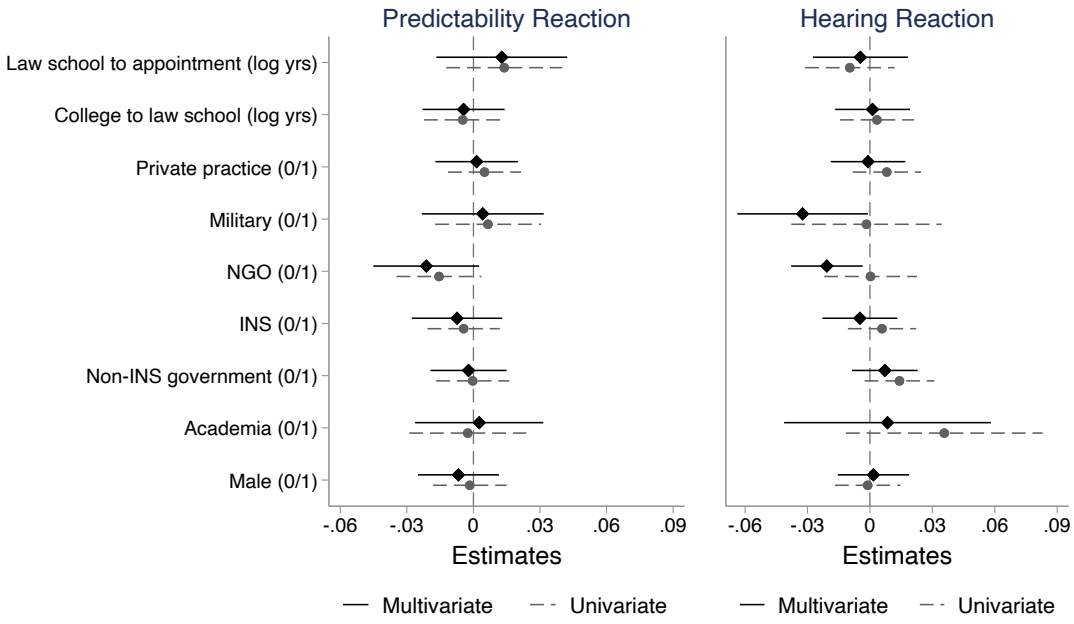
Notes: The figure plots the estimated coefficients $\hat{\mu}_2$ from Specification (5.2), the estimated coefficient of the interaction between the surprise reversal indicator and the period indicator for the first two periods following a surprise reversal. The outcome variable is the daily average number of hearings per proceeding described in Section 4.3. The sample includes all appeal decisions from 1986 to 2019, trimmed using the nearest-event attribution method described in Section 4.1. All estimations are unweighted and include appeal fixed effects and appeal ruling calendar-time fixed effects. Standard errors are robust to heteroskedasticity. Each dot represents an estimate for one of 380 judges, ranked by the magnitude of $\hat{\mu}_2$; gray lines denote 95% confidence intervals.

Figure A.14: Correlations of Judge-Year Level Attentiveness Measures



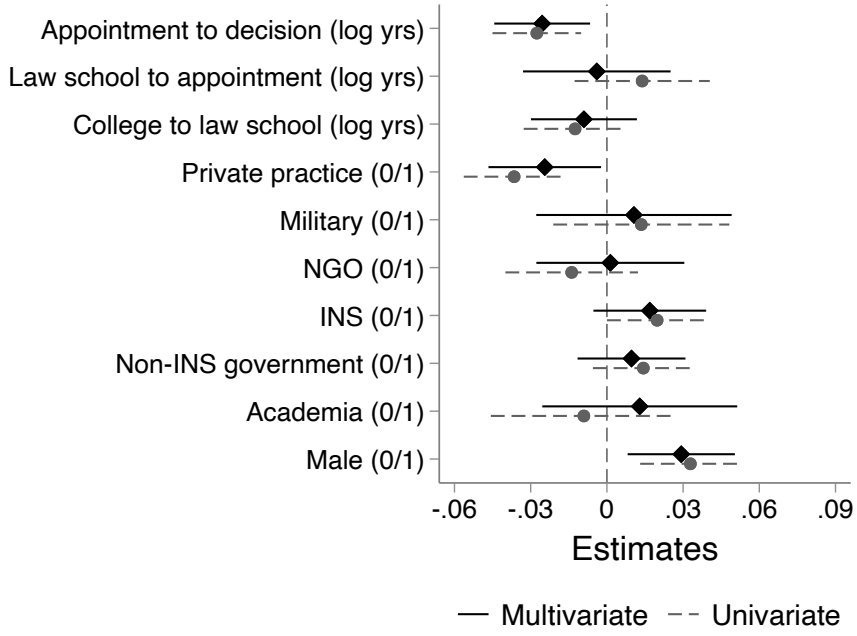
Notes: The figure plots the correlations between the three attentiveness measures described in Section 5.2.1. Each dot represents the attentiveness of a judge in a year in the two dimensions shown on the axes. The size of each dot reflects the number of observations used to generate the estimate for the dimension on the x-axis. The fitted line is weighted by the number of observations used to construct each corresponding estimate, and the gray shaded area indicates the 95% confidence interval around the fitted line.

Figure A.15: Judge Level Early Predictability and Hearing Diligence Reaction and Judge Characteristics



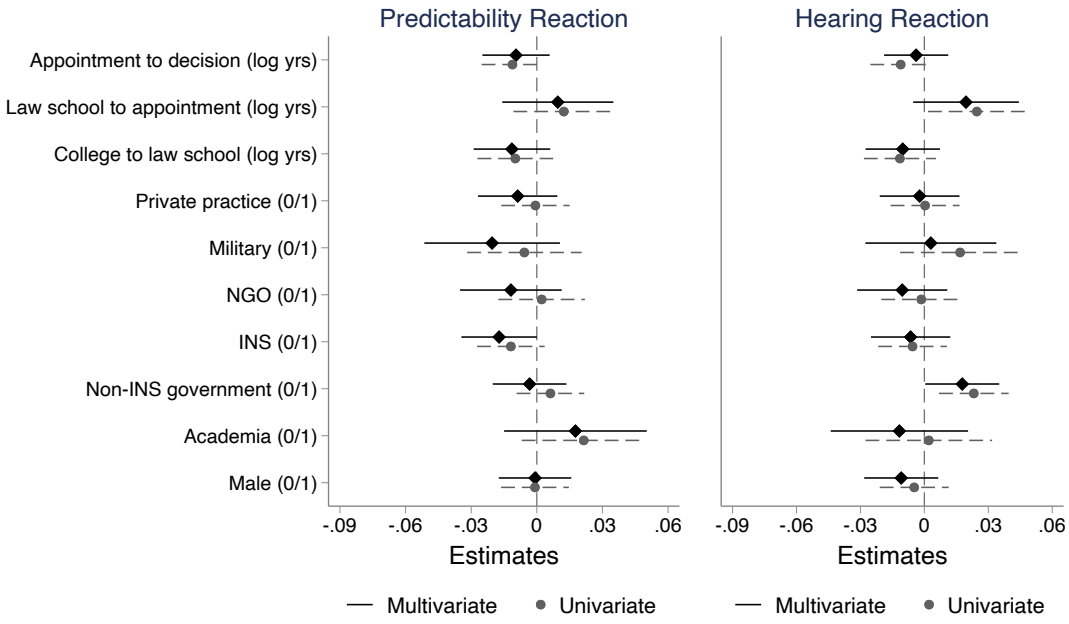
Notes: The figure plots the correlations between judges' biographical characteristics and two attentiveness measures. The left panel shows correlations with the judge-level early predictability reaction, and the right panel shows correlations with the judge-level hearing diligence reaction. Diamonds represent coefficients from multivariate regressions including all characteristics as regressors, with solid lines denoting 95% confidence intervals. Circles represent coefficients from univariate regressions, with dashed lines denoting 95% confidence intervals. All regressions are weighted by the number of proceedings ruled by each judge. Standard errors are robust to heteroskedasticity.

Figure A.16: Judge-Year Level Grant Rate Attentiveness and Judge Characteristics



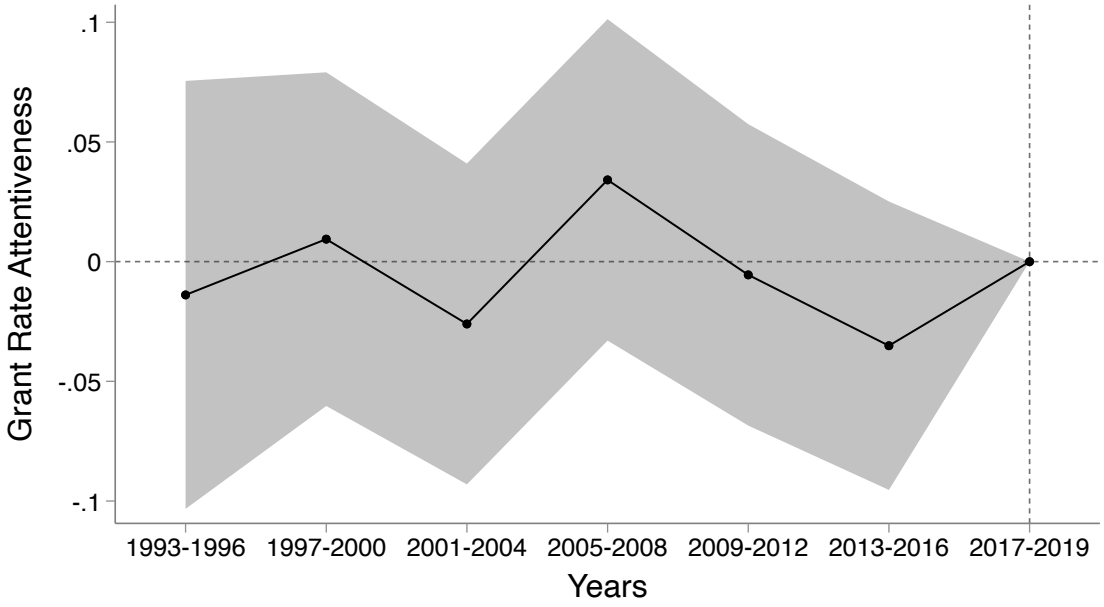
Notes: The figure plots the correlations between the judge-year level grant-rate attentiveness measure and judges' biographical characteristics. Diamonds represent coefficients from multivariate regressions including all characteristics as regressors, with solid lines denoting 95% confidence intervals. Circles represent coefficients from univariate regressions, with dashed lines denoting 95% confidence intervals. All regressions include calendar year fixed effects and are weighted by the number of proceedings ruled by each judge in each year. Standard errors are robust to heteroskedasticity.

Figure A.17: Judge-Year Level Early Predictability and Hearing Diligence Reaction and Judge Characteristics



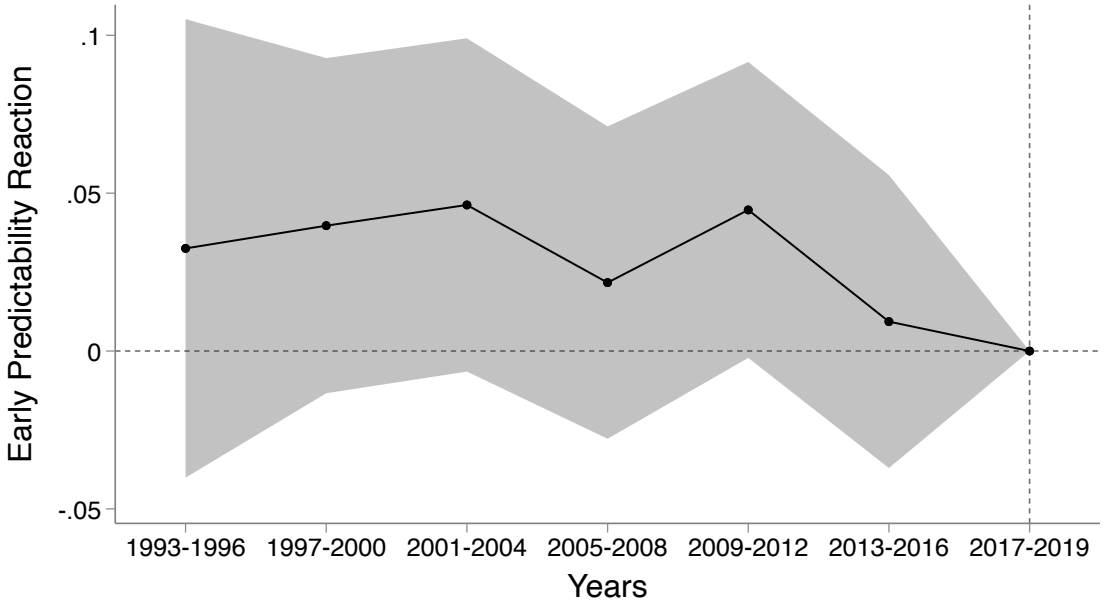
Notes: The figure plots the correlations between judges' biographical characteristics and two attentiveness measures. The left panel shows correlations with the judge-year level early predictability reaction, and the right panel shows correlations with the judge-year level hearing diligence reaction. Diamonds represent coefficients from multivariate regressions including all characteristics as regressors, with solid lines denoting 95% confidence intervals. Circles represent coefficients from univariate regressions, with dashed lines denoting 95% confidence intervals. All regressions include calendar year fixed effects and are weighted by the number of proceedings ruled by each judge. Standard errors are robust to heteroskedasticity.

Figure A.18: Judge-Year Grant Rate Attentiveness Over Presidential Cycles



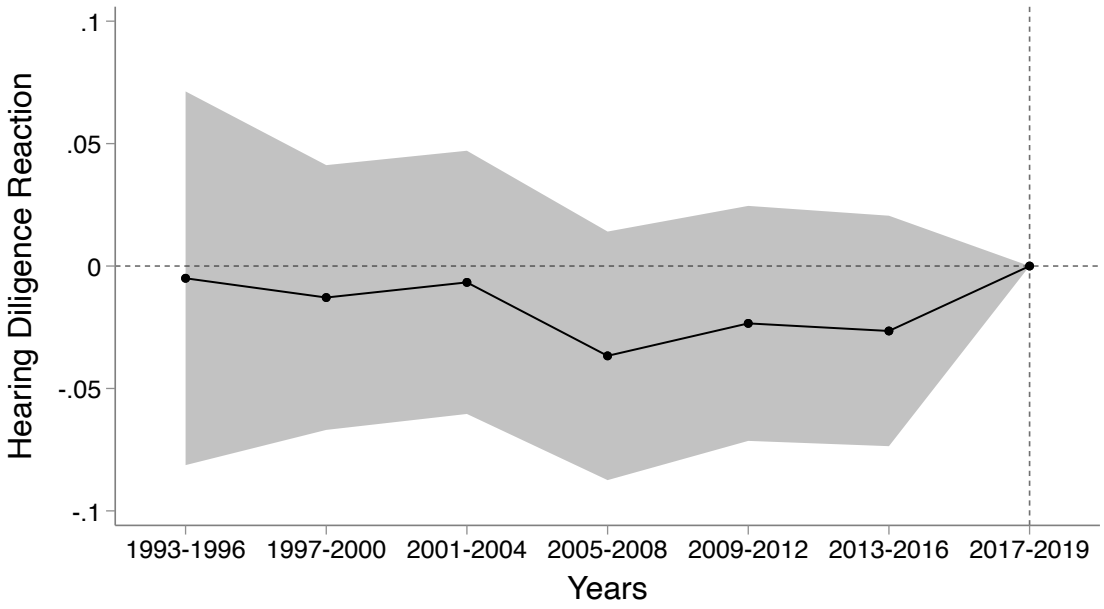
Notes: The figure plots the trend of judges' grant rate attentiveness from 1993 to 2019, using the Trump presidency (2017-2019) as the reference period. Circles represent the estimates for the period indicators, with shaded area representing 95% confidence intervals. All regressions include judge individual fixed effects and are weighted by the number of proceedings ruled by each judge in each year. Standard errors are robust to heteroskedasticity.

Figure A.19: Judge-Year Early Predictability Reaction Over Presidential Cycles



Notes: The figure plots the trend of judges' early predictability reaction from 1993 to 2019, using the Trump presidency (2017-2019) as the reference period. Circles represent the estimates for the period indicators, with shaded area representing 95% confidence intervals. All regressions include judge individual fixed effects and are weighted by the number of proceedings ruled by each judge in each year. Standard errors are robust to heteroskedasticity.

Figure A.20: Judge-Year Hearing Diligence Reaction Over Presidential Cycles



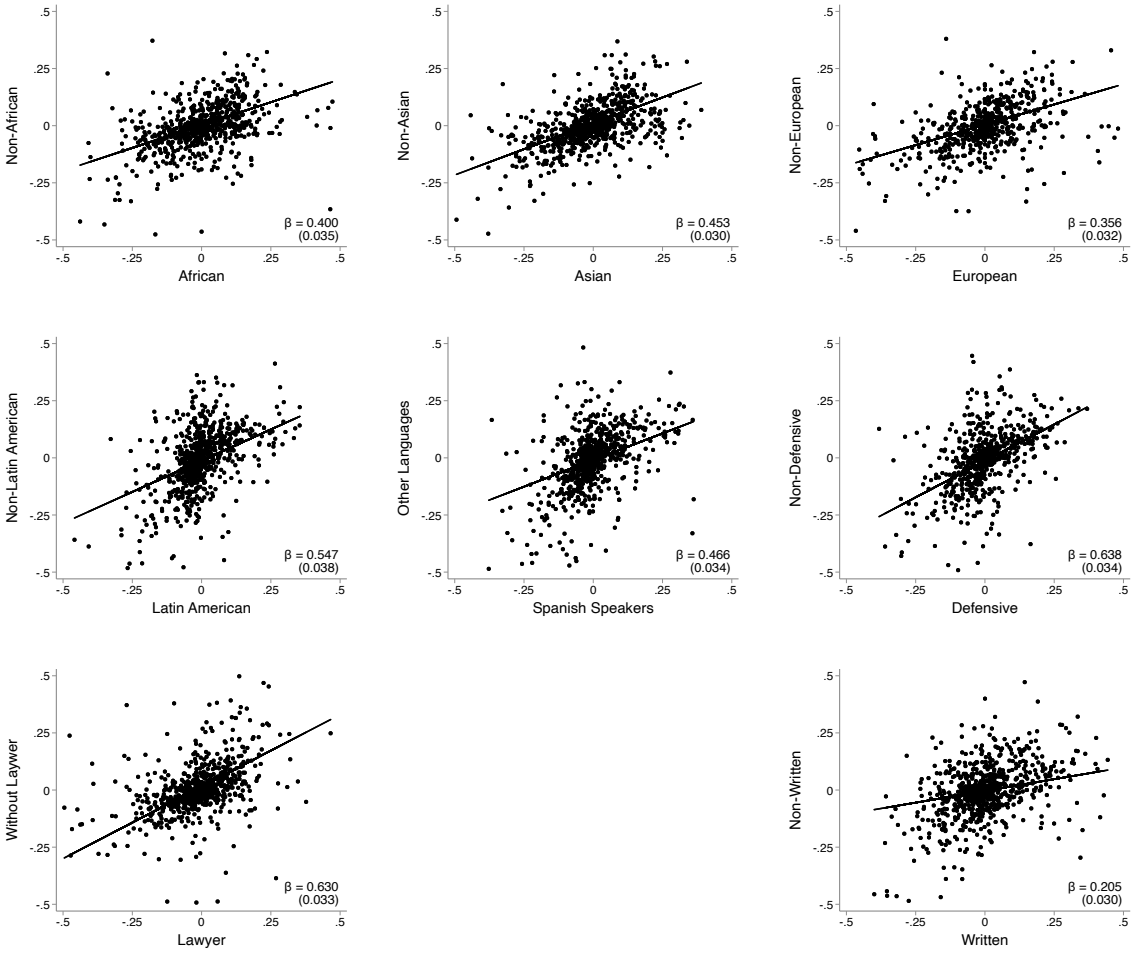
Notes: The figure plots the trend of judges' hearing diligence reaction from 1993 to 2019, using the Trump presidency (2017-2019) as reference period. Circles represent the estimates for the period indicators, with shaded area representing 95% confidence intervals. All regressions include judge individual fixed effects and are weighted by the number of proceedings ruled by each judge in each year. Standard errors are robust to heteroskedasticity.

Table A.8: Randomization Test

	(1)	(2)
With a lawyer	0.077***	0.013***
	(0.004)	(0.002)
Defensive	-0.098***	-0.019***
	(0.005)	(0.003)
Written decisions	-0.089***	-0.019**
	(0.013)	(0.008)
Asian applicants	0.114***	0.029***
	(0.020)	(0.006)
European applicants	0.158***	0.023***
	(0.021)	(0.006)
Latin American applicants	-0.146***	0.002
	(0.021)	(0.006)
African applicants	0.177***	0.020***
	(0.021)	(0.006)
Spanish speaking applicants	0.021**	-0.000
	(0.008)	(0.002)
Order within the day	0.013***	0.001
	(0.001)	(0.001)
Hour of starting	-0.005	-0.005
	(0.015)	(0.013)
Clock time of starinng	0.000	0.000
	(0.000)	(0.000)
In morning	-0.005	-0.014
	(0.018)	(0.018)
During lunchtime	-0.009	-0.019**
	(0.008)	(0.008)
Number of Charges	-0.015***	-0.001
	(0.004)	(0.002)
Number of applications	-0.193***	-0.014
	(0.013)	(0.012)
Days between filing and opening	0.000***	-0.000
	(0.000)	(0.000)
Joint F-test	239.54	10.21
	[0.000]	[0.000]

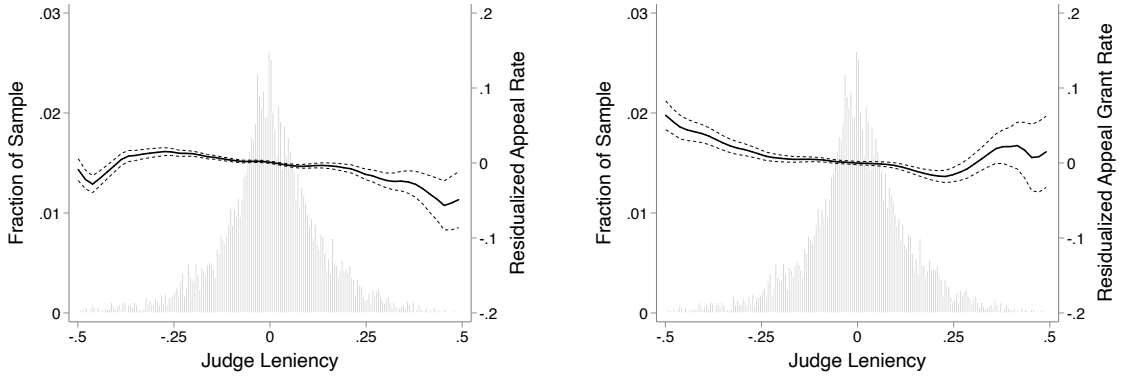
Note: This table reports the randomization test results for the case assignment in the lower court. The sample include all lower court cases from 1984 to 2019. The leave-out judge leniency measure is constructed as described in Section 5.2.2. Column (1) reports the estimates from an OLS regression of lower court grant on the list of variables, controlling for the court-by-time fixed effects. Column (2) reports the estimates from an OLS regression of constructed leniency on the list of variables, controlling for the court-by-time fixed effects. The $F-$statistics and $p-$values of the joint significance test are reported at the bottom. Standard errors clustered at the judge level are reported in parentheses. Asterisks denote significance at the 1% (***), 5% (**), and 10% (*) levels.

Figure A.21: Judge Leniency by Case Characteristics



Notes: The figures show the correlation between the residualized judge leniency measure for different groups of cases. The solid line shows the best linear fit estimated with OLS regressions. Correlation coefficients and standard errors are reported.

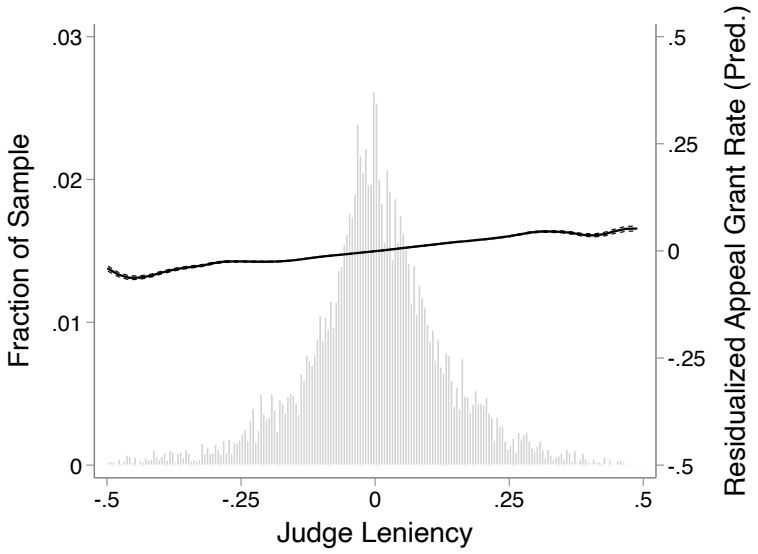
Figure A.22: Leniency and the Two Margins of Reversals



(a) Selection Margin (b) Success Margin

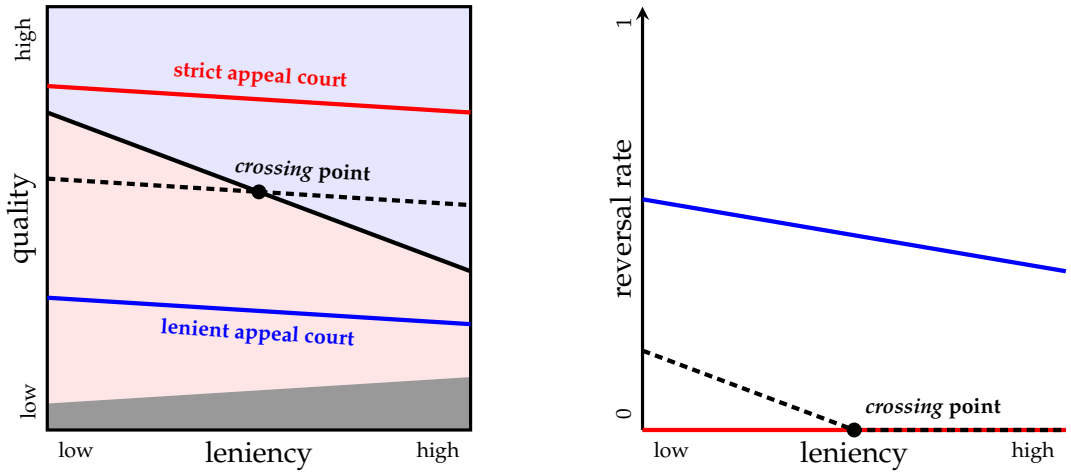
Notes: The figures plot the relationships between judge leniency and the two margins of the reversal rate. In the left panel, we plot the residualized selection margin, appeal rate of denied cases, against leniency. In the right panel, we plot the residualized success margin, appel grant rate of cases that have filed an appeal, against leniency. In both panels, the leave-out judge leniency measure is constructed as described in Section 5.2.2. The solid line shows a local linear regression of the residualized margin on judge leniency. Both margins are residualized using the same set of court-by-time fixed effects applied in constructing the leniency measure in the full sample. Standard errors are clustered at the judge level, and dashed lines denote 95% confidence intervals.

Figure A.23: Judge Leniency and Predicted Appeal Grant Rates



Notes: The figures plot the relationships between judge leniency and the predicted appeal grant rate of all cases in the lower court. The leave-out judge leniency measure is constructed as described in Section 5.2.2. The solid line shows a local linear regression of residualized predicted appeal grant rates on judge leniency. Predicted appeal grant rates are generated with the random forest model trained on the actual appeal decisions described in Section 3.2, and are residualized using the same set of court-by-time fixed effects applied in constructing the leniency measure in the full sample. Standard errors are clustered at the judge level, and dashed lines denote 95% confidence intervals.

Figure A.24: A Stylized Illustration of Leniency and Reversal, Calibrated

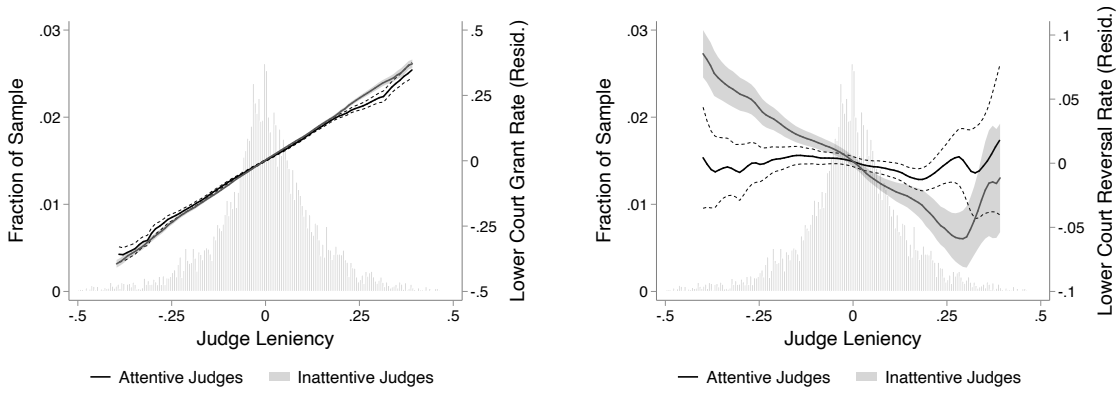


(a) Lower Court Ruling

(b) Reversals

Notes: Panel A.24a illustrates the stylized relationship case quality, judge leniency and lower court decisions across appeal courts of varying stringency, the gray area represents denied cases that do not appeal, they are assumed to be of quality lower than any quality threshold. Panel A.24b displays the corresponding appeal grant rate by leniency for appeal courts of corresponding stringency. Coordinates and slopes are schematic and chosen for expositional clarity.

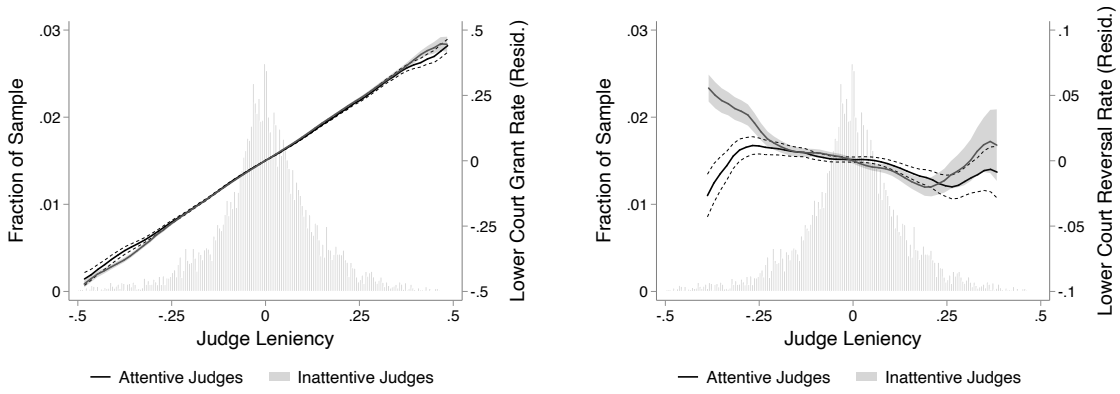
Figure A.25: Leniency and Lower Court Outcomes by Judge-Year Grant Rate Attentiveness



(a) Grant Rate (b) Reversal Rate

Notes: The figures plot the relationships between judge leniency and lower-court outcomes for attentive and inattentive judges, defined by the magnitude of the judge-year effect of surprise reversals on grant rates and classified into the top and bottom terciles, as described in Section 5.3. The leave-out judge leniency measure is constructed as described in Section 5.2.2 for the two subsamples. The left panel shows local linear regressions of residualized lower-court asylum grants on judge leniency, and the right panel shows local linear regressions of residualized lower-court reversals, defined as lower-court denials that are granted on appeal on judge leniency. Both lower-court outcomes are residualized using the same set of court-by-time fixed effects applied in constructing the leniency measure. In both panels, the black line represents the subsample of cases ruled by attentive judges, with dashed lines indicating 95% confidence intervals, while the gray line represents the subsample of cases ruled by inattentive judges, with shaded areas denoting 95% confidence intervals. Standard errors are clustered at the judge level.

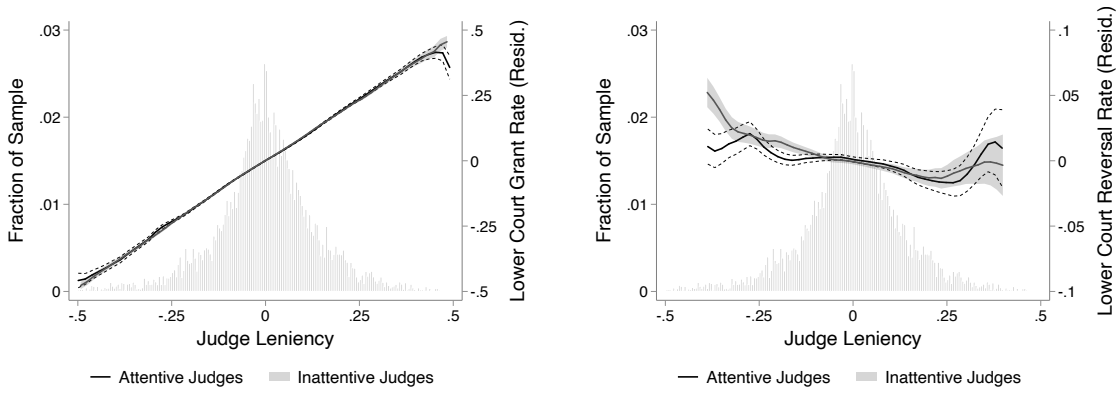
Figure A.26: Leniency and Lower Court Outcomes by Grant Rate Attentiveness, Top Half versus Bottom Half



(a) Grant Rate (b) Reversal Rate

Notes: The figures plot the relationships between judge leniency and lower-court outcomes for attentive and inattentive judges, defined by the magnitude of the effect of surprise reversals on grant rates and classified into the top and bottom halves, as described in Section 5.3. The leave-out judge leniency measure is constructed as described in Section 5.2.2 for the two subsamples. The left panel shows local linear regressions of residualized lower-court asylum grants on judge leniency, and the right panel shows local linear regressions of residualized lower-court reversals, defined as lower-court denials that are granted on appeal on judge leniency. Both lower-court outcomes are residualized using the same set of court-by-time fixed effects applied in constructing the leniency measure. In both panels, the black line represents the subsample of cases ruled by attentive judges, with dashed lines indicating 95% confidence intervals, while the gray line represents the subsample of cases ruled by inattentive judges, with shaded areas denoting 95% confidence intervals. Standard errors are clustered at the judge level.

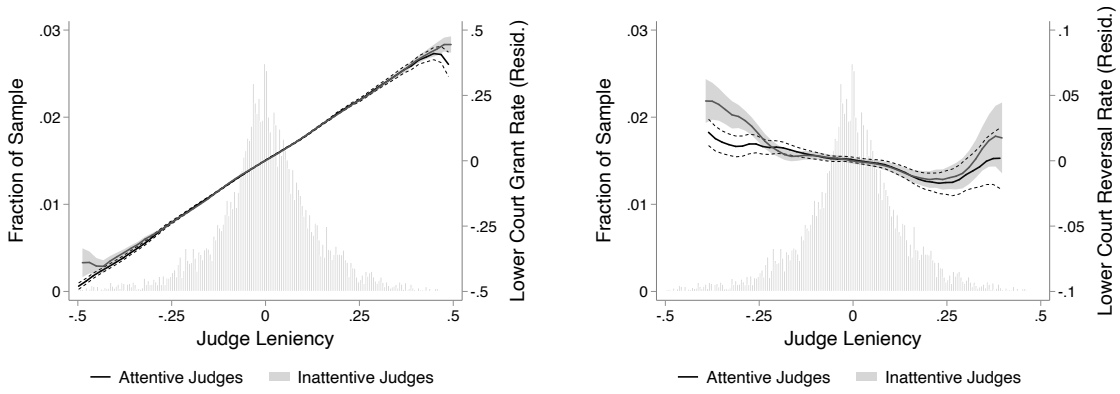
Figure A.27: Leniency and Lower Court Outcomes by Early Predictability Attentiveness



(a) Grant Rate (b) Reversal Rate

Notes: The figures plot the relationships between judge leniency and lower-court outcomes for attentive and inattentive judges, defined by the effect of surprise reversals on early predictability and classified into positive reacting and negative reacting, as described in Section 5.3. The leave-out judge leniency measure is constructed as described in Section 5.2.2 for the two subsamples. The left panel shows local linear regressions of residualized lower-court asylum grants on judge leniency, and the right panel shows local linear regressions of residualized lower-court reversals, defined as lower-court denials that are granted on appeal on judge leniency. Both lower-court outcomes are residualized using the same set of court-by-time fixed effects applied in constructing the leniency measure. In both panels, the black line represents the subsample of cases ruled by positive reacting judges, with dashed lines indicating 95% confidence intervals, while the gray line represents the subsample of cases ruled by negative reacting judges, with shaded areas denoting 95% confidence intervals. Standard errors are clustered at the judge level.

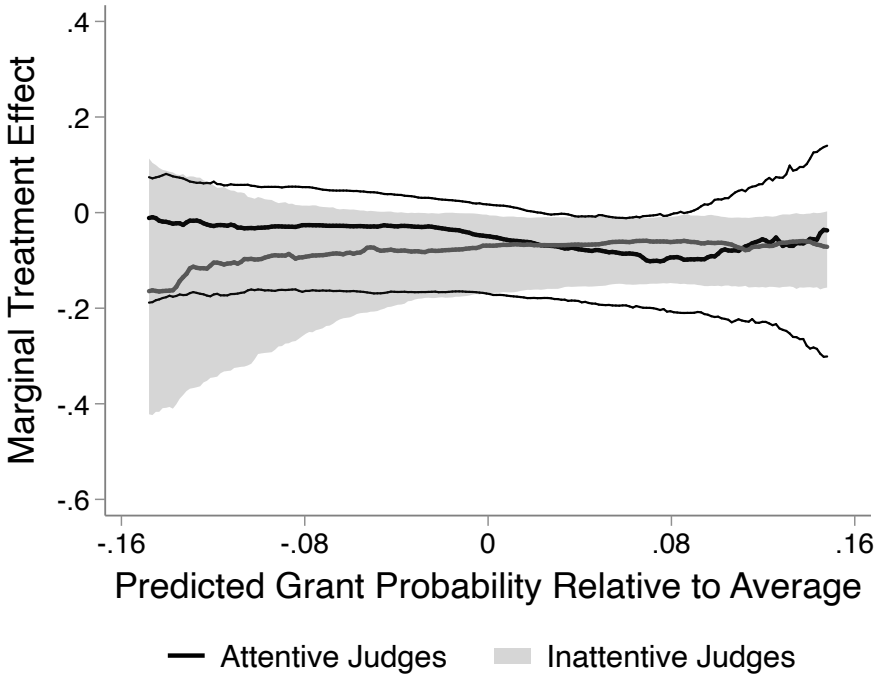
Figure A.28: Leniency and Lower Court Outcomes by No. Hearings per Proceeding Attentiveness



(a) Grant Rate (b) Reversal Rate

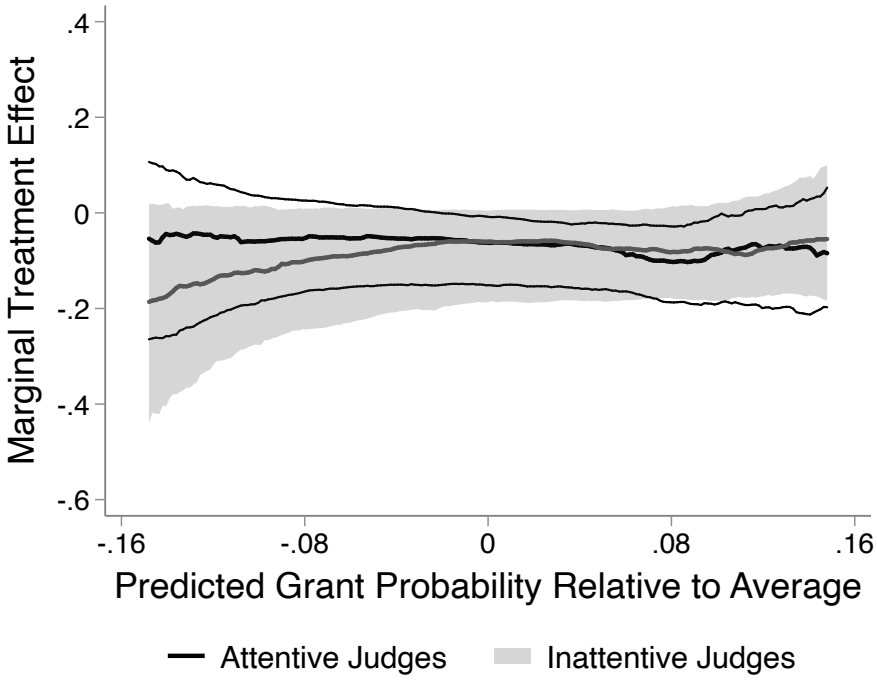
Notes: The figures plot the relationships between judge leniency and lower-court outcomes for attentive and inattentive judges, defined by the effect of surprise reversals on the number of hearings per proceeding and classified into positive reacting and negative reacting, as described in Section 5.3. The leave-out judge leniency measure is constructed as described in Section 5.2.2 for the two subsamples. The left panel shows local linear regressions of residualized lower-court asylum grants on judge leniency, and the right panel shows local linear regressions of residualized lower-court reversals, defined as lower-court denials that are granted on appeal on judge leniency. Both lower-court outcomes are residualized using the same set of court-by-time fixed effects applied in constructing the leniency measure. In both panels, the black line represents the subsample of cases ruled by positive reacting judges, with dashed lines indicating 95% confidence intervals, while the gray line represents the subsample of cases ruled by negative reacting judges, with shaded areas denoting 95% confidence intervals. Standard errors are clustered at the judge level.

Figure A.29: Marginal Treatment Effect, Early Predictability Attentiveness



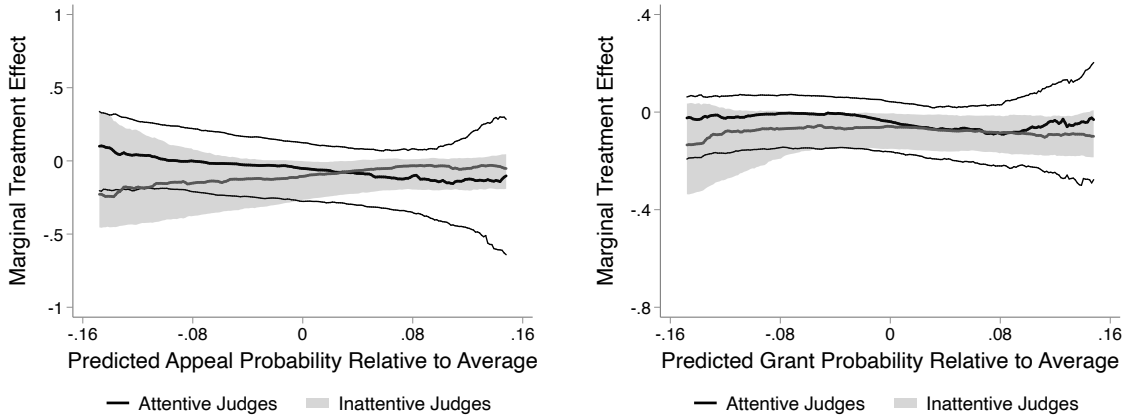
Notes: The figure plots the marginal treatment effects (MTEs) of lower-court decisions on lower-court reversals separately for cases ruled by attentive (negatively reacting) and inattentive (positively reacting) judges, classified by early predictability attentiveness as described in Section 5.2.1. The MTEs are estimated following the procedure outlined in Section 5.3.2. The black solid line shows the MTEs estimated on the subsample of cases ruled by inattentive judges who increase early predictability after surprise reversals, with the dashed lines as 90 percent confidence intervals. The gray solid line shows those estimated on the subsample of cases ruled by attentive judges who decrease early predictability after surprise reversals, with the shaded area as 90 percent confidence intervals. Standard errors are computed using 500 bootstrap replications clustered at the judge level.

Figure A.30: Marginal Treatment Effect, No. Hearings per Proceeding Attentiveness



Notes: The figure plots the marginal treatment effects (MTEs) of lower-court decisions on lower-court reversals separately for cases ruled by attentive (positively reacting) and inattentive (negatively reacting) judges, classified by no. hearings per proceeding attentiveness as described in Section 5.2.1. The MTEs are estimated following the procedure outlined in Section 5.3.2. The black solid line shows the MTEs estimated on the subsample of cases ruled by attentive judges who increase the average number of hearings per proceeding after surprise reversals, with the dashed lines as 90 percent confidence intervals. The gray solid line shows those estimated on the subsample of cases ruled by inattentive judges who decrease the average number of hearings per proceeding after surprise reversals, with the shaded area as 90 percent confidence intervals. Standard errors are computed using 500 bootstrap replications clustered at the judge level.

Figure A.31: Marginal Treatment Effect Decomposition, Early Predictability Attentiveness

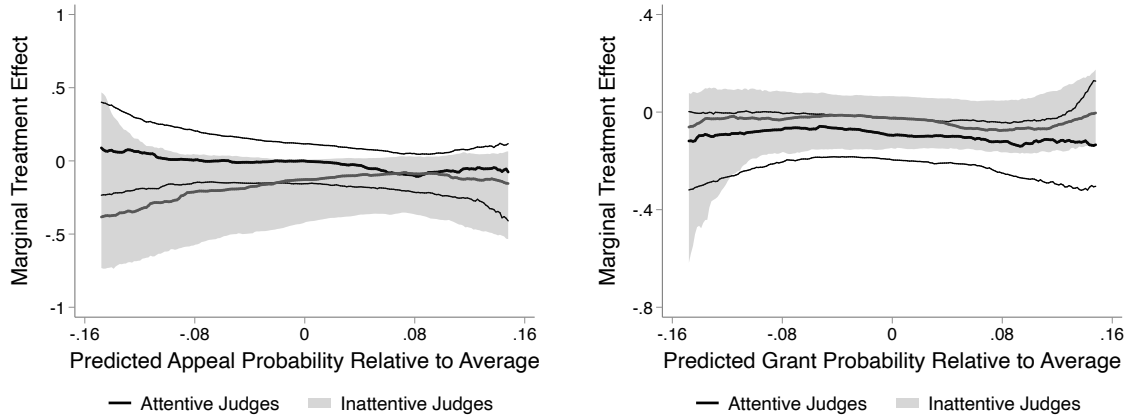


(a) Selection Margin

(b) Success Margin

Notes: The figure plots the marginal treatment effects (MTEs) of lower-court decisions on appeals and appeal success separately for cases ruled by attentive and inattentive judges, classified by early predictability attentiveness as described in Section 5.2.1. The MTEs are estimated following the procedure outlined in Section 5.3.2. The left panel plots the MTEs on appeal filing among all denied cases, the right panel plots the MTEs on appeal grant, conditional on appealing. In both panels, the black solid line shows the MTEs estimated on the subsample of cases ruled by attentive judges, with the dashed lines as 90 percent confidence intervals. The gray solid line shows those estimated on the subsample of cases ruled by inattentive judges, with the shaded area as 90 percent confidence intervals. Standard errors are computed using 500 bootstrap replications clustered at the judge level.

Figure A.32: Marginal Treatment Effect Decomposition, No. Hearing per Proc. Attentiveness



(a) Selection Margin

(b) Success Margin

Notes: The figure plots the marginal treatment effects (MTEs) of lower-court decisions on appeals and appeal success separately for cases ruled by attentive and inattentive judges, classified by number of hearings per proceeding attentiveness as described in Section 5.2.1. The MTEs are estimated following the procedure outlined in Section 5.3.2. The left panel plots the MTEs on appeal filing among all denied cases, the right panel plots the MTEs on appeal grant, conditional on appealing. In both panels, the black solid line shows the MTEs estimated on the subsample of cases ruled by attentive judges, with the dashed lines as 90 percent confidence intervals. The gray solid line shows those estimated on the subsample of cases ruled by inattentive judges, with the shaded area as 90 percent confidence intervals. Standard errors are computed using 500 bootstrap replications clustered at the judge level.

B Details of ML

In this section, we describe the details of the random-forest based prediction model and present additional results for the appeal court analysis. For the lower court prediction exercise, we refer interested readers to the detailed discussion in Dunn et al. (2017).

B.1 Model Training and Implementation

First, we describe the implementation details of the machine-learning framework, including the data preprocessing procedures, the hyperparameter tuning strategy, and the sequential modeling design used for both the lower-court and appeal-court predictions.

B.1.1 Data Preprocessing Pipeline

We implemented a structured data-preprocessing pipeline to prepare the lower-court and appeal-court samples for model estimation. The process consisted of three main steps. First, we imputed missing values using customized functions that allowed variable-specific methods with explicit handling of numerical features that lacked a defined imputation rule. Second, we coerced numerical and categorical variables into their appropriate data types to ensure consistency across the modeling pipeline. Third, we one-hot encoded categorical variables using a fixed set of possible values to guarantee that the training and testing data shared identical feature spaces. After transformation, we split the data into training and testing samples using a reproducible random partition and applied the same preprocessing to both sets. This modular pipeline allowed transparent control over imputation, data-type conversion, and feature encoding, ensuring that all input features were standardized, comparable across datasets, and suitable for downstream random-forest model training.

B.1.2 Hyperparameter Tuning

For the lower-court prediction, we conducted a grid search with three-fold stratified cross-validation. The parameter grid included combinations of maximum tree

depth (30 or unrestricted), the number of estimators (50, 100, 200, or 400), minimum samples required for a split (2 or 5), and minimum samples per leaf (1 or 2). We selected the model that maximizes the ROC-AUC score. The final specification is a random forest with 400 trees and a minimum split size of 5. The out-of-sample ROC-AUC is 0.873 and and the accuracy is 0.809. Due to computational constraints, the tuning was performed on the full sample instead of sequentially. The chosen parameter set was then applied consistently across the smaller early-predictability model variants based on the combined feature set (Judge ID, Nationality, Opening Year-Month, and Filing Year-Month).

For the appeal court prediction, we tuned the random-forest model similarly with a three-fold stratified cross-validation grid search. The parameter grid included maximum tree depth values of 20, 40, and unrestricted; the number of estimators (50, 100, 200, and 400); minimum samples per split (2 or 5); and minimum samples per leaf (1, 2, or 5). The ROC-AUC-maximizing specification, with 400 trees, a minimum split size of 2 and minimum sample leaf of 1, achieved an out-of-sample ROC AUC of 0.854 and an accuracy of 0.797. As with the lower-court model, tuning was conducted on the aggregate sample rather than sequentially. To benchmark model performance, we compared random forests with alternative algorithms, including Extra Trees, Gradient Boosting, Histogram-based Gradient Boosting, and Logistic Regression. The random-forest model consistently outperformed these alternatives as shown in Figure B.1.

B.1.3 Sequential Modelling Approach

To mimic the sequential nature of the decision process, we estimated a series of prediction models separately for each year from 1986 to 2019. For each prediction year, the model is trained only on cases filed in preceding years; for instance, to predict outcomes for cases filed in 1986, the training sample includes all cases filed in 1985 and earlier. As shown in Figure B.2, the appeal grant rate fluctuates considerably over time, suggesting that the predictive relevance of past cases may decay as they become temporally distant. To account for this, we applied an exponential sample-weighting scheme that discounts older observations in the training data.

Formally, let τ denote the number of years between the prediction year and the year of the training observation. Each observation is assigned a weight α given by $\alpha = \delta^{\tau-1}$, where $\delta \in [0, 1]$ controls the decay rate.

We tuned the optimal decay parameter δ^* in two stages. In the first stage, we searched over a coarse grid from 0.1 to 1.0 in increments of 0.1; in the second, we refine the search using a 0.01 grid centered around the best-performing value from the first stage. Model performance is evaluated using the ROC AUC criterion. The optimal decay parameter is 0.90 for the lower-court prediction model and 0.89 for the appeal-court prediction model.

B.2 Appeal Prediction Evaluation

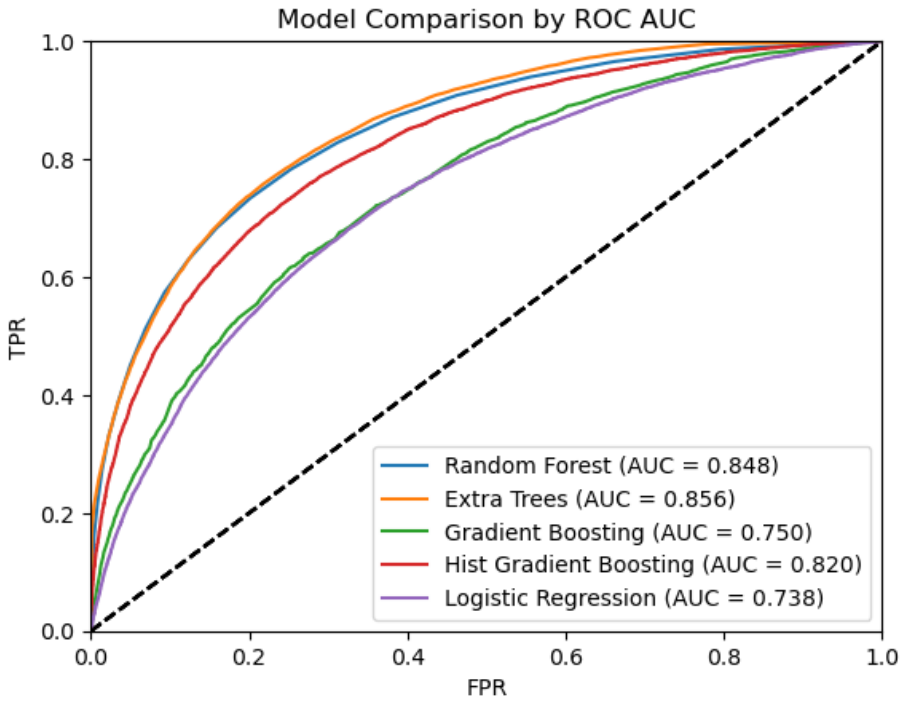
In this section, we further analyze the appeal prediction results. Figure B.3 plots the model's annual accuracy over the sample period. Three main patterns stand out: First, as expected, the sequential modeling approach where only data from preceding years are used for training yields slightly lower accuracy than the aggregate model trained on the full sample. Second, exponential decay weights do not meaningfully improve predictive performance, suggesting that patterns in appeal decision making are relatively stable over time rather than short-lived. Third, while the model performs consistently well overall, it exhibits some year-to-year variation. The most pronounced decline occurs in 2009, largely driven by cases involving Haitian applicants. Following the catastrophic earthquake that struck Haiti in January 2010, many individuals whose asylum applications were denied in 2009 were subsequently granted relief on appeal. As shown in Figure B.4, the temporary drop in model accuracy is almost entirely attributable to this shock and Haitian applicants, underscoring how large, unanticipated geopolitical events can reduce predictive accuracy.

To have a better understanding of the model, we examine the relative importance of each feature across specifications. The variables with the highest predictive power include the time elapsed between the lower-court decision and the appeal filing, applicant nationality, and the lower-court decision date. Nationality is as the most important predictor in the sequential models. In the aggregate model, the algorithm can infer contemporaneous shocks and trend effects by exploiting correlations among appeals filed within the same period. By contrast, the sequential model cannot incorporate information from concurrent or future events.

Lastly, we group all features of the appeal prediction model into six categories, as summarized in Table B.1, and evaluate their relative importances by group. This

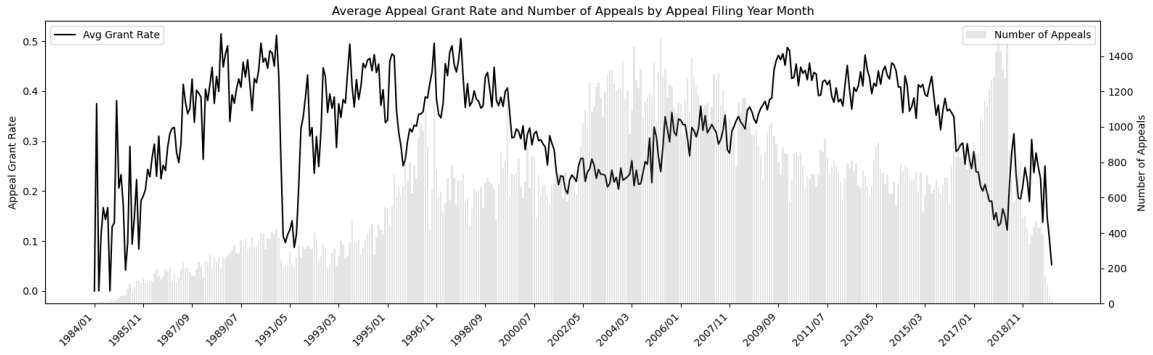
aggregation allows us to quantify the influence of the lower-court judges who initially denied the cases on subsequent appeal outcomes. As shown in the table, features related to lower-court judges account for 0.285 of the total model importance, underscoring the substantial role of judge characteristics in explaining appeal success.

Figure B.1: Appeal Prediction Model Comparison



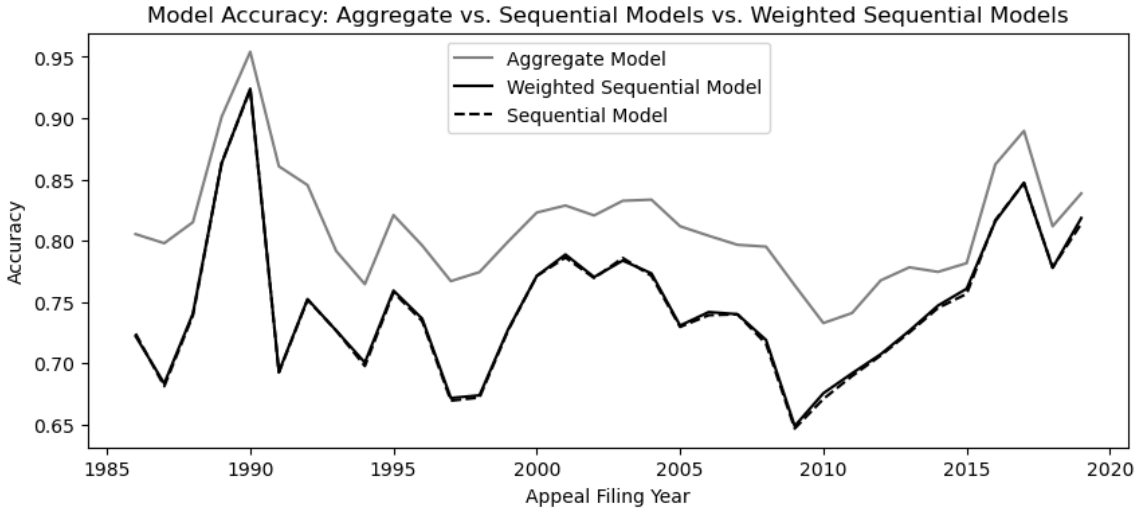
Notes: The figure compares the predictive performance of five classification algorithms for the appeal-court prediction task using receiver operating characteristic (ROC) curves. The models were trained and tested on the aggregate level.

Figure B.2: Monthly Appeal Grant Rate, 1984–2019



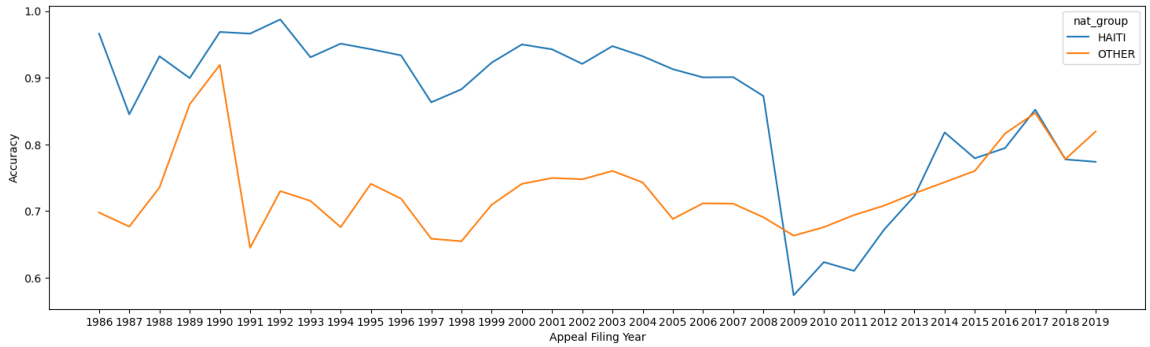
Notes: The figure plots the monthly average appeal grant rate (solid line, left axis) and the number of appeals filed (gray bars, right axis) from 1984 to 2019.

Figure B.3: Prediction Accuracy of the Tuned Models, 1986–2019



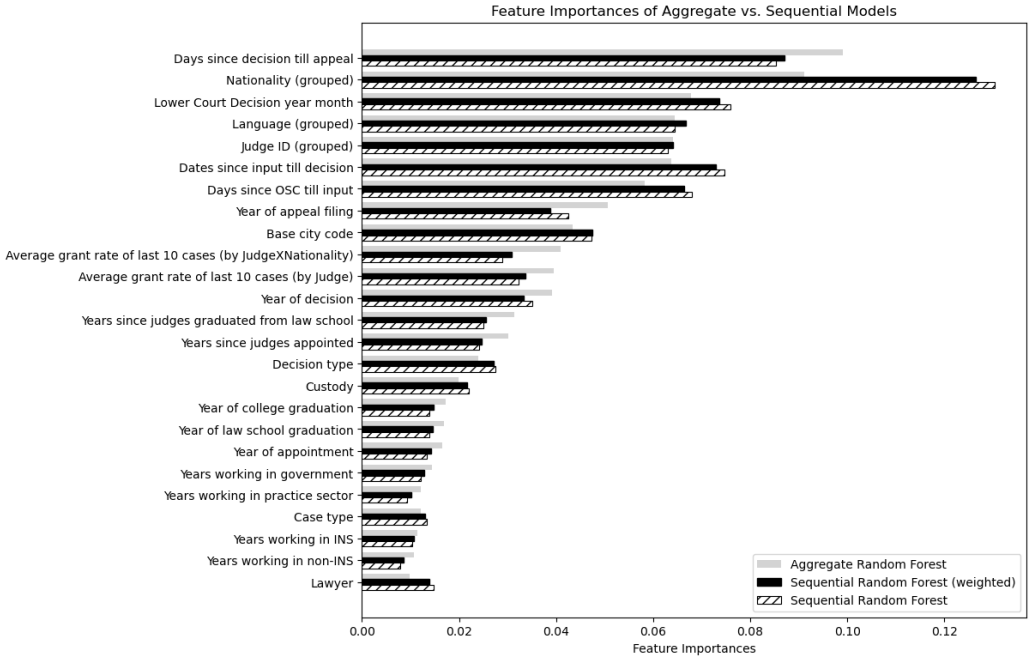
Notes: The figure plots the prediction accuracy of the aggregate model, the unweighted sequential model and the optimally weighted sequential model over the sample period.

Figure B.4: Prediction accuracy, Haiti vs. Others, 1986–2019



Notes: The figure plots the prediction accuracy of the optimally weighted sequential model over the sample period, for applicants from Haiti versus other countries.

Figure B.5: Feature Importances of the Appeal Prediction Model



Notes: The figure compares the relative feature importances of the aggregate model for predicting appeal outcomes, and the mean feature importances of the unweighted sequential model and the weighted sequential model.

Table B.1: Feature Groups and Aggregate Importance

Feature Group	**Feature Importance**	**Features Included**
Time horizon features	0.375	appeal filing year, appeal filing year month, lower court decision year, lower court decision year month, days between lower court filing and case opening, days between lower court case opening and decision, days between lower court decision and appeal filing
Judge features	0.285	judge id, male, year of appointment, year of college graduation, year of law school graduation, years in government, years in non-INS government, years in INS, years in military, years in NGO, years in private practice, years in academia, indicator of missing biography, years since appointment, years since law school graduation, average appeal grant rate of last 10 occasions
Respondent features	0.174	origin nationality, language
Trend features	0.073	average appeal grant rate of last 10 occasions for a judge, average appeal grant rate of last 10 occasions for a judge-nationality group
Proceeding features	0.071	with lawyers, defensive, affirmative, oral, written, case type text, detailed decision text, custody, probono
Location features	0.045	base city, indicator of hearing locations matching with base cities

Notes: The table reports grouped average feature importances derived from the sequential random-forest model. The third column lists all individual features included within each group.

C Proof of Proposition 5.1 and Lemma 5.1

Proof of Proposition 5.1 For $c > c^*$, under the assumption that $f(c) > 0$, $c'(l) < 0$, the reversal rate is defined as

$$R(c) = \int_{c^*}^{c} f(q)\mathrm{d}q.$$

By the Fundamental Theorem of Calculus,

$$\frac{\mathrm{d}R(c)}{\mathrm{d}l} = \frac{\mathrm{d}R(c)}{\mathrm{d}c} \cdot \frac{\mathrm{d}c}{\mathrm{d}l} = f(c) \cdot c'(l) < 0.$$

The selection margin $E(c)$ is defined as

$$E(c) = \frac{F(c) - F(a)}{F(c)},$$

then

$$\begin{aligned}\frac{\mathrm{d}E(c)}{\mathrm{d}l} &= -\frac{\mathrm{d}}{\mathrm{d}c}\left(\frac{F(a)}{F(c)}\right) \cdot \frac{\mathrm{d}c}{\mathrm{d}l} \\ &= -\left(-\frac{F(a)f(c)}{F(c)^2}\right) \cdot c'(l) < 0.\end{aligned}$$

The success margin $I(c)$ is defined as

$$I(c) = \frac{F(c) - F(c^*)}{F(c) - F(a)},$$

then

$$\begin{aligned}\frac{\mathrm{d}I}{\mathrm{d}l} &= \frac{f(c)\left(F(c) - F(a)\right) - \left(F(c) - F(c^*)\right) f(c)}{\left(F(c) - F(a)\right)^2} \cdot \frac{\mathrm{d}c}{\mathrm{d}l} \\ &= \frac{f(c) \cdot \left(F(c^*) - F(a)\right)}{\left(F(c) - F(a)\right)^2} \cdot c'(l) < 0\end{aligned}$$

■

Proof of Lemma 5.1 The sensitivities are defined as function of the threshold c:

$$\psi_R(c) = \frac{\partial R}{\partial c}(c) = f(c),$$
$$\psi_E(c) = \frac{\partial E}{\partial c}(c) = \frac{F(a)f(c)}{F(c)^2}, \qquad \psi_I(c) = \frac{\partial I}{\partial c}(c) = \frac{f(c)\left[F(c^*) - F(a)\right]}{\left[F(c) - F(a)\right]^2}.$$

Assume $f'(c) > 0$ and $f'(c)F(c) > 2f(c)^2$ on $[c^*, c_{j,0}]$, then

$$\frac{\mathrm{d}\psi_R}{\mathrm{d}c}(c) = f'(c) > 0,$$
$$\frac{\mathrm{d}\psi_E}{\mathrm{d}c}(c) = F(a)\frac{f'(c)F(c) - 2f(c)^2}{F(c)^3} > 0,$$

that is, the reversal sensitivity and the selection margin sensitivity functions are both monotonically increasing in c on $[c^*, c_{j,0}]$. Assume $f'(c)\left[F(c) - F(a)\right] > 2f(c)^2$, then

$$\frac{\mathrm{d}\psi_I}{\mathrm{d}c}(c) = \frac{f'(c)\left[F(c) - F(a)\right] - 2f(c)^2}{\left[F(c) - F(a)\right]^3} \cdot \left[F(c^*) - F(a)\right] > 0,$$

that is, the success margin sensitivity function is also monotonically increasing in c on $[c^*, c_{j,0}]$. ■

Assumptions of Lemma 5.1 The regularity conditions we assume for Lemma 5.1, on $[c^*, c_{j,0}]$,

$$f'(c) > 0, \qquad f'(c)F(c) > 2f(c)^2 \qquad f'(c)\left[F(c) - F(a)\right] > 2f(c)^2,$$

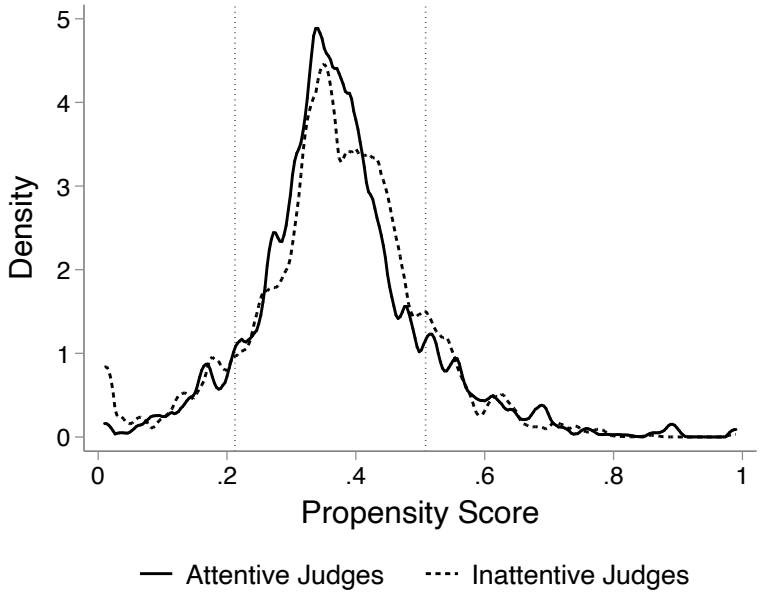
ensure that the sensitivity of reversals and its margins to the grant threshold increases monotonically with judicial strictness over the relevant range.

The first condition requires that the case quality density $f(c)$ increasing over the range of strict thresholds $[c^*, c_{j,0}]$. The second condition requires that the density rise sufficiently quickly relative to the mass of denials, captured by $F(c)$, such that a small movement in a strict threshold reclassifies many denials. The third condition is stricter, requiring that the density rise sufficiently quickly relative to, $F(c) - F(a)$, the mass of appealable denials. Economically, this corresponds to an environment where stricter judges face a thick concentration of borderline cases just below their cutoff and above the appeal-benefit threshold a and are strong enough to succeed

if appealed. Therefore, marginal shifts in the threshold have large effects on reversals, appeal filing, and appeal success. In contrast, when the judge's threshold is already close to the appeal court standard c^*, the pool of borderline strong but denied cases becomes thin and small threshold adjustments have limited effects.

These regularity conditions we impose align with both prior evidence and our own data. Ramji-Nogales, Schoenholtz and Schrag (2007) documented that case difficulty in the U.S. asylum system increases sharly near the higher quality region, where more cases accumulate because they are more likely to be filed, pursued, and appealed. In Figure C.1, we plot the empirical distribution of the estimated grant propensity p as discussed in Section 5.3.2. The distribution is unimodal and right skewed: The density increases drastically from the lower tail and peaks in the interior of the distribution. The interval of relevance $[c^*, c_{j,0}]$ lies in this rising segment of the density, implying that stricter judges operate in the part of the distribution with many appealable denials. Hence, we argue that the regularity conditions are empirically plausible and consistent with the observed distributional shape.

Figure C.1: Distribution of Estimated Propensity Scores



Notes: The figure plots the distribution of the estimated propensity scores $\tilde{p}_{ijt}$ as described in Section 5.3.2. The solid line plots the distribution of cases ruled by attentive judges, the dashed line plots the distribution of cases ruled by inattentive judges. The classification of judges is defined in Section 5.2.1. The vertical lines denote the 10th and 90th percentile of the distribution of all cases.

D Decomposing MTEs

In this section, we derive the MTE decomposition of Equation (5.5). For the sake of simplicity and clarity, we ignore the subscripts for case i ruled by lower court judge j at time t here. Since we define $A \in \{0,1\}$ as the indicator that a denied case files an appeal, and $S \in \{0,1\}$ as indicator that an appealed case is granted on appeal. The reversal rate can then be written as

$$Y_{ijt} \equiv A_{ijt} \cdot S_{ijt},$$

By the law of iterated expectations, we get

$$\begin{aligned} E\left[Y \mid p\right] &= E\left[A \times S \mid p\right] \\ &= E\left[E\left[A \times S \mid A, p\right] \mid p\right] \\ &= E\left[A \times E\left[S \mid A=1, p\right]\right] \\ &= E\left[A \mid p\right] E\left[S \mid A=1, p\right] \end{aligned}$$

Within court-time cells C_{ct},

$$E\left[Y \mid p, C\right] = E\left[A \mid p, C\right] \cdot E\left[S \mid A=1, p, C\right]. \tag{D.1}$$

Let the residualized variable $W \in \{Y, A, S\}$ controlling for court-time fixed effects be

$$\tilde{W} \equiv W - E\left[W \mid C\right],$$

we can then derive the decomposition for the expectation of $\tilde{Y}$ as

$$\begin{aligned} E\left[\tilde{Y} \mid p, C\right] =& E\left[Y \mid p, C\right] - E\left[Y \mid C\right] \\ =& E\left[A \mid p, C\right] E\left[S \mid A=1, p, C\right] - E\left[A \mid C\right] E\left[S \mid A=1, C\right] \\ =& \left(E\left[A \mid p, C\right] - E\left[A \mid C\right]\right) E\left[S \mid A{=}1, p, C\right] \\ &+ E\left[A \mid C\right] \left(E\left[S \mid A{=}1, p, C\right] - E\left[S \mid A{=}1, C\right]\right). \end{aligned}$$

Thus, the residualized mean of Y decomposes into a *selection* deviation and a *success* deviation around their C–specific baselines.

Differentiate D.1 w.r.t. p, we get

$$\frac{\partial}{\partial p} E\left[Y \mid p, C\right] = \underbrace{\frac{\partial}{\partial p} E\left[A \mid p, C\right] \cdot E\left[S \mid A{=}1, p, C\right]}_{\text{selection margin}} + \underbrace{E\left[A \mid p, C\right] \cdot \frac{\partial}{\partial p} E\left[S \mid A{=}1, p, C\right]}_{\text{success margin}}.$$

Since $E\left[W \mid C\right], W \in \{Y, A, S\}$ does not depend on p, the same derivative holds for residuals:

$$\frac{\partial}{\partial p} E\left[\tilde{W} \mid p, C\right] = \frac{\partial}{\partial p} E\left[W \mid p, C\right],$$

that is, residualization does not change the local derivative (the MTE), it only re-centers the levels within court-time cells. We can then decompose the total MTE as

$$\begin{aligned} \text{MTE}_Y(p) \equiv & \frac{\partial}{\partial p} E\left[\tilde{Y} \mid p\right] \\ = & \underbrace{\text{MTE}_A(p)}_{\equiv \frac{\partial}{\partial p} E\left[\tilde{A} | p\right]} \cdot E\left[S \mid A{=}1, p, C\right] + E\left[A \mid p, C\right] \cdot \underbrace{\text{MTE}_S(p)}_{\equiv \frac{\partial}{\partial p} E\left[\tilde{S} | A=1, p\right]}, \end{aligned}$$

where we use $E\left[\cdot \mid C\right]$ and $E\left[\tilde{\cdot}\right]$ interchangeably since the residualization for all variables controls for court-by-time fixed effects.