When Hierarchies Teach: Selective Review and Decision Alignment in U.S. Asylum Courts *

Sai Zhang
USC

Daniel L. Chen
Toulouse

September 5, 2026

[Click for the latest version]

Abstract

Hierarchical review evaluates experts' first-instance decisions and informs their subsequent choices, but this feedback is selectively generated. We study this process in U.S. asylum courts, where feedback follows a denial, applicant appeal, and Board of Immigration Appeals (BIA) disposition. Using linked immigration-court and BIA records from 1986 to 2019, we develop a rolling prediction model that provides an observable ex ante reversal-risk benchmark for classifying realized reversals by signal diagnosticity. Weekly event studies show a 1.57-percentage-point larger first-week increase in lower-court grant rates after surprise than after expected reversals, with no broad decline in observable hearing-diligence proxies. We develop a threshold model of response magnitude and distributional alignment with the BIA standard. High-response judges have flatter marginal treatment effects (MTEs) than low-response judges, especially at below-average predicted grant probabilities. The pattern appears on both appeal filing among denials and BIA success conditional on appeal, with the clearest contrast on appeal filing. This alignment gap is smaller in defensive than affirmative proceedings, consistent with structured hearings making feedback more usable. Review's teaching function therefore depends on feedback selection, signal diagnosticity, and recipient responsiveness.

*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. We also thank participants at the ALEA 2026 annual meeting. An earlier version of this paper circulated under the title *Judicial Learning from Feedback: Evidence from U.S. Asylum Courts*. Zhang: University of Southern California, saizhang@usc.edu. Chen: Toulouse School of Economics, daniel.chen@iast.fr.

1 Introduction

Organizations delegate consequential decisions to experts because they hold local information that is costly to centralize. Delegation preserves the value of this information but creates a fundamental control problem: how can a hierarchy align decentralized decisions with a higher-level standard? Existing theories of hierarchy emphasize how organizational architecture allocates decision rights, preserves incentives for local information use, and filters decision errors (Sah and Stiglitz, 1986; Aghion and Tirole, 1997). A related literature on performance feedback shows that hierarchy also shapes how evaluative signals are interpreted and translated into subsequent decisions (Joseph, Klingebiel and Wilson, 2016). Studying this teaching function is often difficult because review produces selective and incomplete feedback. Experts' own decisions determine what becomes eligible for review, while downstream choices determine what ultimately reaches the reviewing authority. Even feedback with the same formal direction can differ in how much it reveals about the higher-level standard. This paper asks when selectively generated feedback changes future decisions and whether stronger responsiveness is associated with closer alignment with that standard.

We study this question in a distinctive setting: the U.S. immigration court system. It combines decentralized, high-stakes adjudication with review by the Board of Immigration Appeals (BIA), a common national authority. In the channel we study, formal feedback is selectively generated through an immigration judge's denial, the applicant's decision to appeal, and a BIA disposition. Because immigration judges make repeated decisions over time, we can link appellate feedback to their subsequent behavior within the same hierarchy. To study this process, we assemble four data sources linked through proceeding, case, and immigration-judge identifiers. The core EOIR data contain 832,946 first-instance asylum proceedings decided between 1986 and 2019, which we combine with detailed hearing histories, BIA appeal records, and hand-collected judge biographies. These linked records allow us to follow lower-court decisions into appellate review and then observe the same judges' subsequent rulings.

Our first set of results uses two rolling random-forest models with distinct purposes. The lower-court model predicts final rulings from information fixed before the hearing process and achieves 76.7% out-of-sample accuracy, providing

an early-predictability benchmark used in the later analysis. The BIA model predicts appeal grants and denials from pre-disposition observables with 74.8% out-of-sample accuracy. Because these inputs are observable to immigration judges, the BIA model provides a disciplined benchmark for the expectations judges could feasibly form about their own reversal risk. Among realized reversals, we classify model-predicted appeal grants as expected reversals and model-predicted appeal denials as surprise reversals. The two groups share the same formal appellate direction, but a surprise reversal carries a sharper informational message: the BIA grants an appeal that was difficult to anticipate from the information available before resolution. Relative to an expected reversal, it therefore provides a more diagnostic signal that the judge's decision standard differs from the BIA standard.

Second, we use weekly event studies to compare changes in the subsequent grant rates of immigration judges around surprise and expected reversals. From the recorded BIA decision date through the following six days, grant rates rise by 1.57 percentage points more after a surprise reversal than after an expected reversal, relative to the seven days immediately preceding the decision. Because both events reverse a denial but differ in model-based diagnosticity, this differential response shows that immigration judges respond not only to the formal direction of appellate review but also to the information it conveys. We next examine two families of observable proxies for lower-court hearing diligence: early predictability and procedural measures constructed from hearing records. The stronger grant-rate response is not accompanied by a broad decline in hearing diligence across either family of proxies. The number of hearing records per proceeding rises modestly in the first post-event week, while the other procedural measures show no systematic immediate change. Together, the grant-rate and process evidence points to a substantive adjustment in lower-court decision-making after more diagnostic appellate feedback.

The event-study evidence raises a deeper question: are the judges who respond most strongly to appellate feedback also more closely aligned with the BIA standard? We address this question with a simple threshold model of attention and learning in which a judge-specific effective updating rate governs how strongly appellate feedback moves the judge's decision threshold toward the BIA standard. Guided by the model, we construct judge-specific measures of changes in grant rates, early predictability, and hearing counts after surprise rather than ex-

pected reversals. We use observed responsiveness as the empirical counterpart of attentiveness in the model. Grant-rate responsiveness is our primary measure because it is based on realized lower-court outcomes and directly captures how judges change their subsequent rulings after appellate feedback. We classify judges with larger and smaller response magnitudes as high-response and low-response judges, respectively.

We combine the response classifications with a leave-out measure of judge leniency, constructed from each judge's grant rate in other cases after accounting for the assignment of cases across judges within courts and time periods. Using the resulting variation in judge leniency, we estimate marginal treatment effects (MTEs) for the probability that a lower-court denial is followed by a BIA grant (Goldsmith-Pinkham, Hull and Kolesár, 2026; Heckman and Vytlacil, 2005; Heckman, Urzua and Vytlacil, 2006). The MTE measures the local change in this denial-conditioned reversal rate as predicted lower-court grant probability rises. Intuitively, a slightly more lenient judge grants some claims that a stricter judge would deny, changing the composition of the remaining denial pool. A negative MTE means that the reversal rate among these denials falls with leniency, while an MTE near zero means that the rate changes little as leniency increases.

For high-response judges, the MTE remains close to zero across the predicted grant-probability distribution. For low-response judges, it is more negative over much of the distribution, with the clearest separation at below-average probabilities, where decision thresholds are stricter. The flatter MTE for high-response judges means that their denial-conditioned reversal rate is less sensitive to differences in first-instance leniency. The model interprets this pattern as closer distributional alignment with the BIA standard. Positive- and negative-reacting high-response judges have MTE curves that are not systematically distinguishable, further supporting response magnitude as the empirical counterpart of attentiveness.

We next examine the two gates of selectively generated review: the appeal-filing margin among lower-court denials and the BIA-success margin among appealed denials. For the appeal-filing margin, the MTE for high-response judges remains closer to zero, while the MTE for low-response judges is negative across the distribution and most negative where decision thresholds are stricter. The BIA-success margin shows the same qualitative ordering, but the separation is smaller and less precisely estimated. Together, the two margins trace the attentiveness-

alignment relationship through both stages of selected review, with the clearest empirical contrast at appeal filing.

The relationship between attentiveness and alignment also differs by procedural setting. For defensive proceedings, MTEs remain comparatively flat even among low-response judges. For affirmative proceedings, the MTE is sharply negative among low-response judges, particularly where decision thresholds are stricter, while remaining comparatively flat among high-response judges. This contrast is consistent with the higher stakes, more developed records, and more structured hearings in defensive proceedings making BIA feedback more salient and easier to apply across similar cases. The larger alignment gap in affirmative proceedings therefore suggests that attentiveness matters most when appellate feedback is less readily transferable across cases.

In addition, source-country comparisons provide supplementary evidence that alignment also varies with substantive case characteristics. The point estimates indicate closer distributional alignment for Latin American origins and weaker alignment for recently conflict-affected origins, while a country-income split produces little separation. The absence of an income gradient shows that not every country classification generates a distinct MTE pattern. Across these comparisons, high-response judges generally exhibit MTEs closer to zero than low-response judges, reinforcing the link between attentiveness and alignment with the BIA standard.

Taken together, our results show how appellate review becomes organizational feedback. More diagnostic reversals generate stronger behavioral responses, and judges who respond more strongly to these signals are more closely aligned with the BIA standard. Across both gates of selectively generated review and different procedural settings, the evidence points to the same organizational principle: the consequences of feedback depend on both the information it contains and the attentiveness of the recipient. Our analysis therefore contributes to organizational economics by treating the review hierarchy as an information technology. The design of review and the effectiveness of hierarchy shape not only how lower-level decisions are evaluated, but also which signals are generated and how recipients incorporate them into future decentralized choices.

Contribution to literature. We make our primary contribution to organizational economics by showing how selective review generates feedback for decentralized experts. Classic work on authority and communication explains how delegation preserves the use of expert information while limiting centralized control (Aghion and Tirole, 1997; Dessein, 2002). Models of knowledge hierarchies and organizational architecture show how organizations route difficult decisions and filter judgment errors (Garicano, 2000; Sah and Stiglitz, 1986). Complaint-based monitoring makes oversight endogenous to affected parties' responses (Prendergast, 2003). Theories of appeals emphasize error correction, incentives, and strategic case selection (Shavell, 1995, 2006; Cameron, Segal and Songer, 2000). Beim (2017) further examines judicial information flowing upward when higher courts learn from lower-court decisions. We examine the feedback generated in the other direction, when selected higher-level review can inform future decisions by lower-level experts. In the applicant-appeal channel, the immigration judge's ruling and the applicant's appeal jointly determine which cases can produce formal BIA feedback. By following this process, we connect feedback production and signal diagnosticity to recipient responsiveness and distributional alignment with the higher-level standard.

The recipient side of selective review connects our study to research on organizational learning from performance feedback. Classic work describes learning as the transformation of experience into routines and organizational knowledge, with knowledge creation, retention, and transfer occurring at different rates (Levitt and March, 1988; Argote and Miron-Spektor, 2011; Argote, Lee and Park, 2021). Experience is itself selective, and organizations can learn differently from successes and failures (Denrell and March, 2001; Madsen and Desai, 2010). How feedback is interpreted and acted upon also depends on organizational structure, while delayed outcomes complicate the connection between decisions and their consequences (Joseph, Klingebiel and Wilson, 2016; Rahmandad, 2008). At the recipient level, learning depends on how attention is allocated and whether decision-makers notice relevant regularities in their experience (Ocasio, 1997; Schwartzstein, 2014; Hanna, Mullainathan and Schwartzstein, 2014). We connect these insights by separating two sources of adjustment after formally similar appellate outcomes: the diagnostic information conveyed by the signal and the magnitude of the recipient's behavioral response.

We also contribute to research on judicial responses to appellate review, including work on fear of reversal, compliance with higher-court directives, and lower-court discretion (Klein and Hume, 2003; Gennaioli and Shleifer, 2008; Westerland et al., 2010). The closest empirical paper is Bhuller and Sigstad (2025), who use quasi-random assignment to appellate panels in Norway to identify criminal judges' responses to sentence-increasing and sentence-reducing reversals. We study a complementary organizational object: feedback is selectively produced through denial, applicant appeal, and BIA disposition, while the focal reversals share a common formal direction but differ in diagnosticity. Judge-specific response magnitude captures recipient heterogeneity, and the MTE evidence relates that heterogeneity to distributional alignment across both the appeal-filing and BIA-success margins and across substantive and procedural settings. The measurement problem is related to selective labels, where prior human decisions determine which outcomes researchers observe (Lakkaraju et al., 2017; Kleinberg et al., 2018). Selective feedback instead concerns which formal evaluations the review process generates for the decision-maker. Using information available before BIA disposition, the rolling prediction model distinguishes expected and surprise reversals within a common formal direction, producing a case-specific measure of signal diagnosticity. This measurement allows us to study when selectively generated review feedback is translated into behavioral responsiveness and distributional alignment with the BIA standard.

Structure of the Paper. The remainder of the paper is organized as follows. Section 2 describes hierarchical review in U.S. asylum courts and the data used in our analysis. Section 3 develops rolling prediction models to measure early ruling predictability and classify appellate feedback. Section 4 presents the event-study design and estimates immigration judges' responses to expected and surprise reversals. Section 5 develops the threshold model of attention and learning and examines how behavioral responsiveness relates to distributional alignment with the BIA standard. Section 6 concludes.

2 Hierarchical Review in U.S. Asylum Courts

This section describes the institutional setting of U.S. asylum courts, the data used in our analysis, and the review hierarchy in which immigration judges make first-instance decisions. We focus on how denial, applicant appeal, and BIA disposition determine which cases generate formal appellate feedback.

2.1 Delegated Asylum Adjudication

Under the 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] Individuals may apply for asylum through the U.S. Citizenship and Immigration Services (USCIS) Asylum Office or in removal proceedings. If USCIS does not grant an application, or if removal proceedings are initiated, the applicant may pursue the asylum claim before an immigration judge within the Executive Office for Immigration Review (EOIR).[2] Within the review hierarchy studied in this paper, immigration judges make first-instance decisions.

A single immigration judge presides over hearings, determines removability, and adjudicates applications for asylum, withholding of removal, or protection under the Convention Against Torture. The immigration judge evaluates testimony and supporting evidence under the Immigration and Nationality Act (INA).

Cases are administratively assigned to immigration judges within each local immigration court.[3] EOIR guidance states that, upon filing, "the case is randomly assigned to a specific judge within that court," with subsequent motions generally decided by the assigned judge absent unavailability.[4] This structure delegates responsibility for each case to an individual immigration judge. The assignment process is also important for our empirical analysis because asylum grant rates

[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.

[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.

vary across immigration judges.[5] Immigration judges also manage substantial caseloads.[6] This workload can create a tension between deciding cases promptly and devoting more time to each adjudication. These constraints shape the environment in which immigration judges exercise discretion.

2.2 Appellate Review and Feedback

In the applicant-appeal channel studied in this paper, appellate review begins when an immigration judge denies asylum and the applicant files a notice of appeal with the Board of Immigration Appeals (BIA). A denial that is not appealed does not generate a formal BIA disposition through this channel, while a grant generally falls outside the applicant-initiated denial-appeal process. The BIA exercises nationwide appellate jurisdiction.[7] The applicant must file the notice of appeal directly with the BIA within the prescribed deadline. It may not be filed with the Department of Homeland Security (DHS) or the Immigration Court. Once an appeal reaches the BIA, the BIA reviews the immigration judge's decision and issues a disposition. Appeals 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.[8] The BIA reviews factual findings under a clear-error standard and reviews questions of law, discretion, or judgment *de novo*.[9]

Appellate review serves two functions in this hierarchy: it evaluates the immigration judge's decision in the appealed case, and the resulting BIA disposition can inform judges' future decisions. We focus on the second function while accounting for the denial and applicant-appeal decisions that determine which cases generate formal BIA feedback.[10]

[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] The median judge in our sample rules on 59 proceedings per year.

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

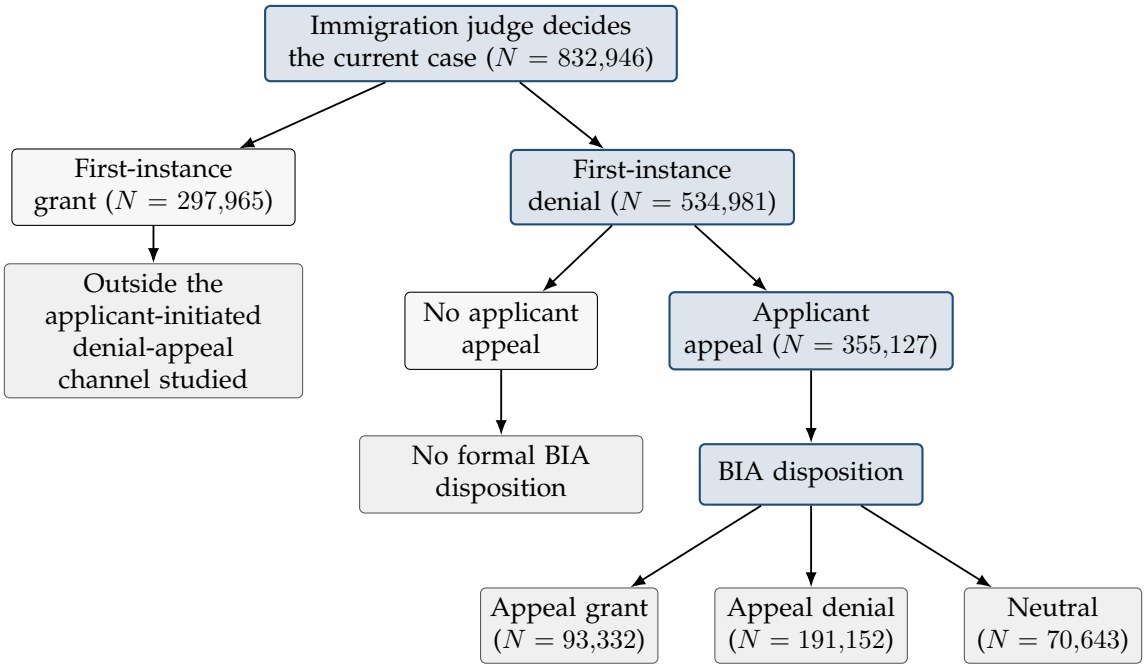
[8] BIA Practice Manual (Composition of the Board). Retrieved from https://www.justice.gov/eoir/reference-materials/bia.

[9] 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.

[10] The applicant-appeal channel is not the only form of review in the immigration system: BIA decisions may also be reviewed by regional U.S. courts of appeals through petitions for review.

The review system gives immigration judges asymmetric exposure to appellate review. A stricter judge denies more cases and therefore creates more opportunities for applicant appeals, but applicants determine which eligible denials reach the BIA. Formal BIA feedback is thus selected by both the immigration judge's initial decision and the applicant's response. This difference in exposure can affect incentives without requiring assumptions about personal sanctions or career consequences. Together with the caseload constraints described above, the selected review environment shapes the information available for future decisions. Figure 1 summarizes the applicant-initiated denial-appeal channel and the corresponding sample counts.

Figure 1: Hierarchical Review as Selective Organizational Feedback



Notes: The figure depicts the applicant-initiated denial-appeal channel studied in the paper. First-instance counts refer to proceedings in the cleaned EOIR sample. Appellate counts refer to BIA appeal records matched to lower-court denials. In the figure, appeal grant, appeal denial, and neutral denote the positive, negative, and neutral-or-other BIA disposition categories defined in Appendix Table A.3. Accordingly, an appeal grant may include a remand, while an appeal denial may include an affirmance or dismissal. A first-instance grant generally lies outside this channel, and a denial without an applicant appeal produces no formal BIA disposition through it. A BIA disposition reviews the current appealed case and may inform future decisions.

For more details of the asylum adjudication process, we refer the interested reader to Ramji-Nogales, Schoenholtz and Schrag (2007). The data described next allow us to observe each stage of this selected review process and to link BIA dis-

positions to immigration judges' subsequent decisions.

2.3 Data

Our analysis combines four data sources linked through proceeding, case, and immigration-judge identifiers. EOIR proceeding records provide lower-court decisions and case characteristics; hearing records provide procedural measures; BIA appeal records provide appeal filings, dispositions, and decision dates; and hand-collected biographies provide observed judge characteristics. These linked data support distinct samples for the prediction, event-study, and MTE analyses, which we identify separately below.

EOIR proceeding data Following previous work (Chen, Moskowitz and Shue, 2016; Dunn et al., 2017), we use EOIR administrative ruling records on asylum proceedings in U.S. immigration courts from 1986 to 2019. After preprocessing and cleaning, the cleaned EOIR proceeding sample contains 832,946 unique proceedings. Of these proceedings, 297,965 (35.8%) resulted in an asylum grant and 534,981 in a denial. For each proceeding, the data report application filing, case opening, and decision dates; the adjudication outcome; court location; and other case and administrative characteristics. Appendix Tables A.1 and A.2 report the ten most common applicant countries of origin and the ten hearing states with the largest numbers of proceedings, respectively.

Hearing data We complement the proceeding data with a hearing-level dataset containing 24,725,971 hearing records from 1984 to 2019. The broader hearing dataset also covers non-asylum, procedural, and administrative matters. Of these records, 4,665,402 link to proceedings in the cleaned EOIR proceeding sample of 832,946. The linked records report hearing dates and clock times and the recorded reasons for adjournments. We use these records to construct the procedural measures examined in the event-study analysis.

BIA appeal data The raw BIA dataset contains 1,108,046 records covering all appeals filed through 2019. We first exclude records that lack the proceeding or case identifier needed for linkage, retaining 89.5% of the raw records. We next restrict

the data to five appeal types used in the analysis: *Appeal of IJ MTR*, *Case Appeal*, *Circuit Court Remand*, *Interlocutory Appeal*, and *MTR BIA*. This leaves 478,006 records. When a proceeding has multiple retained appeal records, we keep the most recent record, leaving 357,766 records. We then exclude 52 records without a valid appeal filing year or case type. We refer to the resulting records as the filtered and deduplicated BIA record sample.

Following the BIA decision classifications summarized in Appendix Table A.3, we group the decisions into three categories: positive, negative, and neutral or other. Positive dispositions include grants, remands, and the other outcomes assigned to this category; negative dispositions include denials, affirmances, dismissals, and the other outcomes assigned to that category. We treat neutral or other dispositions as a separate category without assigning them a common substantive interpretation. Restricting the filtered and deduplicated BIA records to appeals with a positive or negative disposition yields the binary BIA prediction sample of 286,440 appeal records. In later sections, we refer to positive and negative dispositions as appeal grants and appeal denials, respectively.[11] Appeals with neutral or other dispositions do not enter this sample. Of the 534,981 lower-court denials in the cleaned EOIR sample, we match 355,127 to a BIA appeal record. Among these matched appeal records, 93,332 are classified as appeal grants and 191,152 as appeal denials.[12] Appendix Figure A.1 displays the distribution of lower-court decisions and BIA dispositions from 1986 to 2019.

Immigration judge biography data We hand-collected one biography record for each of 698 immigration judges and link these records to EOIR proceedings using judge identifiers. The biographies report appointment year, educational background, and prior professional experience. We use these variables to describe how estimated responses vary with observed judge characteristics. Appendix Table A.4 summarizes selected characteristics of these judges.

The empirical analyses use different subsets of these linked data. Section 3 constructs separate lower-court and BIA prediction samples; the latter is the 286,440-

[11] These terms are shorthand for the disposition categories in Appendix Table A.3. An appeal grant may include a remand or another positive BIA disposition and therefore does not necessarily imply a final grant of asylum. An appeal denial may include an affirmance, dismissal, or another negative BIA disposition and does not necessarily imply a denial on the merits.

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

record binary BIA prediction sample defined above. Section 4 uses BIA decisions from 1986 to 2019 as feedback events and links them to lower-court proceedings decided by the same immigration judge. Analyses of procedural outcomes further require matched hearing records. In Section 5, the appeal-filing MTE uses lower-court denials, while the BIA-success MTE further restricts the sample to appealed denials. The corresponding sections describe the remaining exercise-specific restrictions.

3 Ruling Predictability and Feedback Classification

This section develops two rolling prediction exercises with distinct roles in the empirical analysis. The lower-court model measures how predictable first-instance rulings are from predetermined observables and generates out-of-sample prediction measures used later to examine observable decision-process responses. The BIA model classifies realized appellate feedback by its diagnostic content.

We use random-forest models as flexible prediction tools.[13] In the data, most inputs are categorical, and their relationships with adjudication outcomes may be nonlinear and interactive. Random forests accommodate these features without imposing a linear functional form or requiring a prespecified set of interactions (Breiman, 2001). Appendix Section B.1 reports the preprocessing and tuning procedures for both prediction models in detail.

For both the lower-court and BIA prediction exercises, filing year defines the rolling prediction cohorts. For a cohort filed in year t, we fit the random-forest model using observations filed before year t and apply the fitted model to observations filed in year t. This filing-cohort ordering excludes observations filed in year t or later from that year's model fit, ensuring that model estimation relies exclusively on information available at the time of decision making. Hyperparameters are selected separately using the aggregate-sample tuning procedure described in Appendix Section B.1.

[13] Random forests perform strongly in binary-classification settings with tabular data and often outperform commonly used alternatives (Fernández-Delgado et al., 2014; Grinsztajn, Oyallon and Varoquaux, 2022).

3.1 Early Predictability of Lower-Court Decisions

For the lower-court prediction exercise, our primary objective is to test how *early* immigration judges' decisions can be predicted, that is, how predictable final rulings are from observables fixed before the hearing process starts. The models use judge identifier, applicant nationality, filing timing, and case-opening timing; they do not use hearing or post-hearing information. The prediction target is the immigration judge's final grant-versus-denial decision. The resulting out-of-sample predictions provide the early-predictability measures used in Section 4.3.

To quantify the extent of early predictability, we estimate five specifications using alternative combinations of predetermined feature groups. We construct the lower-court prediction sample from the 832,946 proceedings decided by immigration judges between 1986 and 2019. We exclude proceedings filed before 1984 because of lower data quality in the earlier filing cohorts, leaving 831,656 proceedings. Lower-court prediction cohorts span filing years 1986 through 2019. Extending Dunn et al. (2017), we re-estimate the model annually using proceedings filed in earlier years and predict each cohort out of sample. The five specifications use: (1) judge identifier only; (2) judge identifier and applicant nationality; (3) judge identifier and case-opening date; (4) judge identifier, applicant nationality, and case-opening year-month; and (5) judge identifier, applicant nationality, case-opening year-month, and filing year-month.

For the subsequent analysis, we use specification (5), the most inclusive specification. As shown in Table A.5, this specification achieves an out-of-sample classification accuracy of 76.7% and the lowest log loss among the five specifications. These results document substantial early predictability of final lower-court rulings from observables fixed before the hearing process starts. 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 use the resulting out-of-sample predictions as a benchmark for the ruling that could be anticipated before the hearing process. Greater correspondence between this benchmark and the final ruling indicates a smaller informational contribution from the hearing process to the observed decision. Section 4.3 examines whether this observable decision-process margin changes after appellate feedback.

3.2 Classifying Appellate Feedback

The BIA prediction exercise provides a benchmark for classifying the diagnostic content of appellate feedback. Using information available before the BIA disposition, it distinguishes relatively expected reversals from diagnostically surprising reversals. The classifier uses six groups of observables: timing, immigration-judge, applicant, proceeding, location, and recent appellate-outcome features. Most case, judge, and lower-court timing features are fixed by the lower-court ruling, while appeal-filing timing is realized after that ruling but before the BIA disposition.[14] Together, these variables provide a rich information set that could inform an immigration judge's expectation of reversal risk before BIA resolution. The tuned random-forest model maps this information into a predicted appeal-grant probability, providing a disciplined benchmark for the beliefs judges could feasibly form about their own reversal risk. Appendix Table B.1 lists the input variables.

The model predicts whether a BIA appeal record is classified as an appeal grant or an appeal denial, using the shorthand for the positive and negative disposition categories defined in Section 2.3 and Appendix Table A.3. The resulting binary BIA prediction sample contains 286,440 filtered and deduplicated appeal records. Records with neutral or other dispositions do not enter this sample. For the BIA exercise, appeal filing year defines rolling cohorts from 1986 through 2019. For each cohort, we likewise train the model on appeals filed in earlier years and predict the cohort out of sample.

Across the rolling prediction cohorts, the preferred optimally weighted sequential random-forest model achieves an overall out-of-sample classification accuracy of 74.8%. It performs particularly well in identifying appeal denials, but is less effective in predicting appeal grants: its true negative rate is 90.0%, while its true positive rate is 43.6%.[15] For the event-study analysis in Section 4, we focus on real-

[14] Recent appellate-outcome features include the average appeal-grant rate over the ten preceding appeals, calculated separately for each immigration judge and each judge-by-applicant-nationality group.

[15] The true negative rate is the share of realized appeal denials classified as appeal denials. The true positive rate is the share of realized appeal grants classified as appeal grants. We also apply the BIA classifier to lower-court denials without a matched appeal. Among these nonappealed denials, 7.1% are classified as appeal grants. Among 72,300 lower-court denials classified as appeal grants, 82.8% have a matched appeal filing. These comparisons use the model benchmark to describe selection into observed appeals; nonappealed denials have no realized BIA disposition in the channel studied.

ized reversals, which are lower-court denials followed by appeal grants. We classify a realized reversal as an expected reversal when the model predicts an appeal grant and as a surprise reversal when the model predicts an appeal denial. Both groups have the same formal appellate direction, but surprise reversals were less predictable from pre-resolution observables and therefore carry greater diagnostic information relative to the model-based benchmark. Appendix Section B.2 reports additional evidence on annual predictive performance and feature importance.

4 Immigration Judges' Responses to Appellate Feedback

Section 3.2 classifies realized reversals according to their diagnostic content. This section uses weekly event studies to trace how lower-court judges' subsequent decisions change around BIA dispositions and whether their responses vary with signal diagnosticity.

We estimate two related event studies. The general BIA-outcome specification compares lower-court decision paths around appeal grants and appeal denials. The feedback-diagnosticity specification restricts the event sample to realized reversals and compares surprise reversals with expected reversals based on the classifier in Section 3.2. This within-reversal comparison is our primary specification for examining responses to signal diagnosticity.

4.1 Event Study Design

An event e is a filtered and deduplicated BIA appeal record with an appeal-grant or appeal-denial disposition and a valid recorded BIA decision date, t_e, linked to the immigration judge who issued the corresponding lower-court denial. The general BIA-outcome specification includes both disposition categories, while the diagnosticity specification restricts the event sample to expected and surprise realized reversals. Event time is indexed to t_e, which provides the consistently observed timing anchor in the administrative data.

For each event e, we construct 13 seven-day periods indexed by $k \in \{-6, \ldots, 6\}$, comprising six pre-event periods, the event period, and six post-event periods.

Period k contains lower-court cases decided by the same immigration judge on dates in $[t_e + 7k, t_e + 7k + 6]$. Thus, $k = 0$ begins on the recorded BIA decision date and ends six days later, while $k = -1$ covers the seven days immediately preceding that date and serves as the omitted period. Linking each lower-court case to its eligible event windows generates a dataset of case-event-period observations.

Appendix Figure A.3 displays recurring seven-day variation in the timing of recorded BIA decisions relative to lower-court rulings. We therefore use seven-day periods as a common event-time unit. This aggregation accommodates the recurring timing pattern, stabilizes the probability of treatment across cases and over time, and reduces mechanical day-of-week variation that could otherwise complicate the parallel-trends comparison.

Sample trimming. When an immigration judge has several BIA events close in time, the same lower-court case can enter multiple event windows. A case may then appear before one event and after another. Our preferred **nearest-event attribution** rule retains all linkages for cases that appear exclusively on one side of their eligible events. For cases that appear on both sides, it retains the linkage with the smallest absolute weekly event-time index, $|k|$.[16] This rule reduces conflicting pre-event and post-event assignments while preserving cases that enter only pre-event or only post-event windows.

For robustness, we also use **pre-event-only deduplication**. For cases that enter both pre-event and post-event windows, this rule retains every post-event linkage and only the pre-event linkage with the smallest absolute event-time index, $|k|$.[17] Cases that appear exclusively on one side remain unchanged. This alternative therefore retains more observations than nearest-event attribution.

Nearest-event attribution retains 1,760,863 case-event-period observations, or 16% of the untrimmed case-event-period sample. Pre-event-only deduplication retains 7,293,200 case-event-period observations, or 66.3% of the same sample. Appendix Figure A.4 illustrates the two trimming rules schematically.

[16] If several eligible event linkages share the same minimum value of $|k|$, the construction retains all tied linkages.

[17] If several pre-event linkages share the minimum value of $|k|$, the construction retains all tied linkages.

Event-time binning and event weighting. Appendix Figure A.5 shows that the retained case-event-period observations in the feedback-diagnosticity sample are concentrated near the BIA event under nearest-event attribution. Our main specification therefore pools periods $k = -6, \ldots, -3$ into an early pre-event bin and periods $k = 3, \ldots, 6$ into a late post-event bin, while retaining separate indicators for $k = -2, -1, 0, 1, 2$. Period $k = -1$ remains the omitted category.

Even after trimming, a lower-court case may still be linked to multiple events when all retained linkages place it on the same side of the corresponding BIA events. The unweighted specification gives equal weight to each retained case-event-period observation. Events also differ in the number of lower-court cases observed at a given event time, so events with more linked cases contribute more to the estimated coefficients. If response dynamics vary with event size, the unweighted estimates will reflect this composition. To assess the sensitivity of our results to differences in event size, we also estimate an event-equal weighted specification.

Specifically, for each retained case-event-period observation with lower-court case i to BIA event e at event time k, we assign the weight

$$w_{ie}(k) = \frac{1/N_{e,k}}{\sum_{e'} 1/N_{e',k}} E_k$$

, where $N_{e,k}$ is the number of retained lower-court observations linked to event e at event time k, and E_k is the number of distinct events represented at k. The denominator sums the inverse cell sizes across all events represented in that period. The inverse cell-size component gives each represented event equal total weight, while the remaining terms normalize the event-specific observation weights within period k.

Appendix Figure A.5 compares the raw number of retained case-event-period observations, the number of represented BIA events, and the effective sample size under event-equal weighting.[18] The effective sample size (ESS) equals the raw

[18] Specifically, for each event time k, we calculate

$$ESS_k = \frac{\left(\sum_{i,e} w_{ie}(k)\right)^2}{\sum_{i,e} w_{ie}(k)^2},$$

where the sums extend over all retained case-event-period observations at k.

observation count when all observations receive equal weight and decreases as the weights become more concentrated. Its flatter profile indicates that event-equal weighting makes effective support more even across event time.

We report unweighted estimates as our main specification because they characterize decision paths across the retained case-event-period sample. Following Baker et al. (2026), the robustness specification gives each represented BIA event equal total weight within each event-time period.

Specification and estimation. We first estimate the general BIA-outcome specification:

$$Y_{ijek} = \sum_{\kappa \in K} \left[\alpha_\kappa \mathbf{1}(k \in \kappa) + \beta_\kappa \mathbf{1}(k \in \kappa) \cdot \mathbf{1}\left(\text{Appeal Grant}_e = 1\right) \right] + \mathbf{X}_e' \boldsymbol{\phi} + \varepsilon_{ijek}, \tag{4.1}$$

where Y_{ijek} equals one if lower-court case i, decided by immigration judge j, is granted and zero otherwise. The indicator $\mathbf{1}\left(\text{Appeal Grant}_e = 1\right)$ equals one if BIA event e is an appeal grant and zero if it is an appeal denial. The set

$$K = \{[-6, -3], \{-2\}, \{0\}, \{1\}, \{2\}, [3, 6]\}$$

contains the included event-time bins, with period $k = -1$ omitted. The coefficients α_κ trace the lower-court decision path around appeal-denial events relative to the omitted period. The coefficients β_κ measure how the corresponding path around appeal-grant events differs from the path around appeal-denial events, relative to the same omitted period.

Our primary feedback-diagnosticity specification examines whether immigration judges respond differently to reversals that vary in diagnostic content. Both event groups are realized reversals, defined as lower-court denials followed by appeal grants. The classifier identifies an expected reversal when it predicts an appeal grant and a surprise reversal when it predicts an appeal denial. Because surprise reversals were less predictable from pre-resolution observables, they carry greater

diagnostic content relative to the model-based benchmark. We estimate

$$Y_{ijek} = \sum_{\kappa \in K} \left[a_\kappa \mathbf{1}(k \in \kappa) + b_\kappa \mathbf{1}(k \in \kappa) \cdot \mathbf{1}\left(\text{Surprise Reversal}_e = 1\right) \right] + \mathbf{X}_e' \boldsymbol{\varphi} + \varepsilon_{ijek}, \tag{4.2}$$

where $\mathbf{1}\left(\text{Surprise Reversal}_e = 1\right)$ equals one when event e is an appeal grant that the classifier predicted would be denied and zero when it is an appeal grant that the classifier predicted would be granted. The coefficients a_κ trace the lower-court decision path around expected reversals relative to period $k = -1$. The coefficients b_κ measure how the path around surprise reversals differs from the path around expected reversals, relative to the same omitted period. This within-reversal comparison provides the primary evidence on responsiveness to signal diagnosticity.

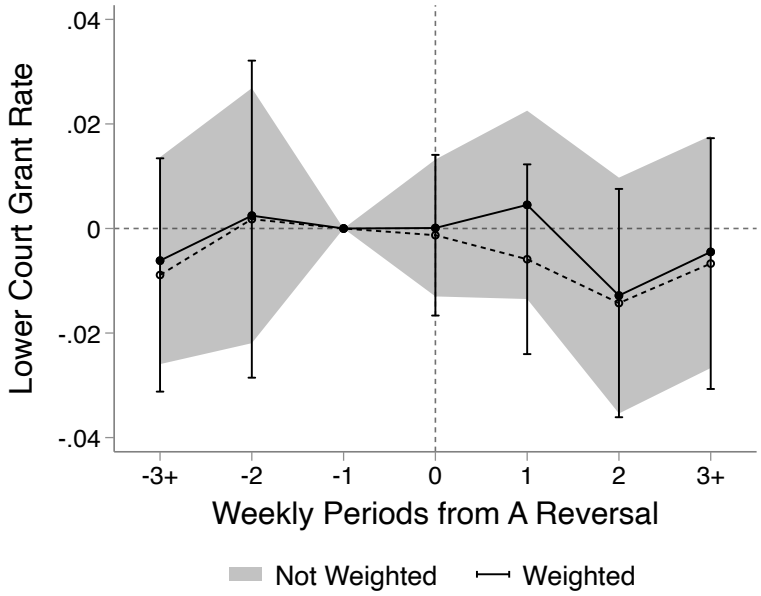
Both specifications include BIA event fixed effects and two sets of calendar fixed effects constructed from the recorded BIA decision date: BIA decision year interacted with day of the week and BIA decision month of the year interacted with day of the week. These fixed effects are represented by $\mathbf{X}_e$ in Equations (4.1) and (4.2). Standard errors are clustered by immigration judge, allowing for dependence across retained observations associated with the same judge.

4.2 Grant-Rate Responses to Appellate Feedback

Figure 2 presents the general BIA-outcome specification, which compares changes in lower-court grant rates around appeal grants with changes around appeal denials. The displayed pre-event coefficients are close to zero and do not show a systematic differential pattern. The post-event coefficients are small and imprecisely estimated over the displayed periods, so the benchmark comparison does not reveal a systematic differential grant-rate response. Appendix Figure A.6 reports the corresponding unbinned estimates under both the unweighted and event-equal-weighted specifications.

The benchmark comparison combines appeal grants that differ in diagnostic content. We therefore turn to our primary feedback-diagnosticity specification, which holds the appellate direction fixed and compares grant-rate responses after surprise reversals with responses after expected reversals. Figure 3 shows that immigration judges' grant rates rise more after surprise reversals than after expected

Figure 2: Lower-Court Grant-Rate Responses: Appeal Grants versus Appeal Denials

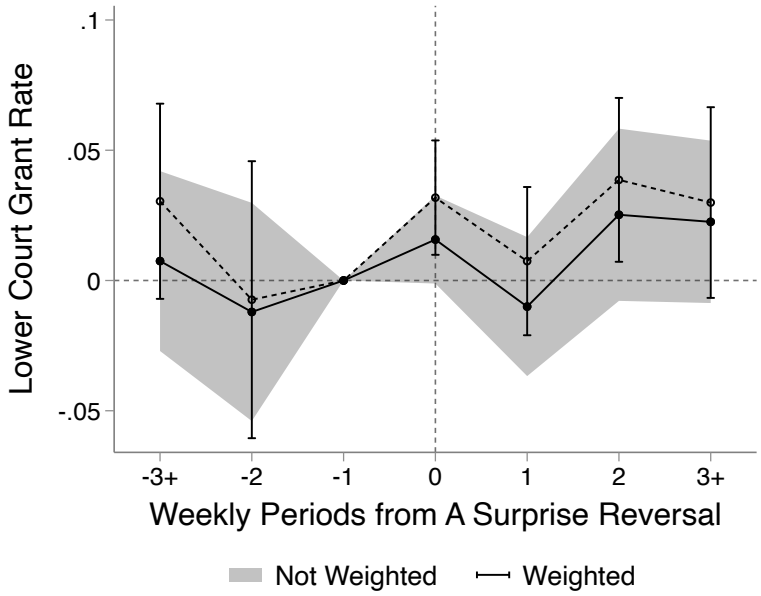


Notes: The figure plots $\hat{\beta}_\kappa$ from Specification (4.1). Each coefficient measures how the change in lower-court grant rates around appeal-grant events differs from the corresponding change around appeal-denial events, relative to the omitted period $k = -1$. The general BIA-outcome event sample contains filtered and deduplicated BIA appeal records with an appeal-grant or appeal-denial disposition, a valid recorded BIA decision date, and a link to the immigration judge who issued the corresponding lower-court denial. Lower-court cases are linked to events using the nearest-event-attribution rule described in Section 4.1. The x-axis indexes binned seven-day periods around the recorded BIA decision date; period $k = 0$ begins on that date and ends six days later. All specifications include BIA event fixed effects, BIA decision-year-by-day-of-week fixed effects, and BIA decision-month-by-day-of-week fixed effects. Standard errors are clustered by immigration judge. Solid dots with shaded pointwise 95% confidence intervals report the preferred unweighted estimates; open circles with capped pointwise 95% confidence intervals report the event-equal-weighted estimates.

reversals. In the preferred unweighted specification, the $k = 0$ estimate is 1.57 percentage points: relative to the seven days immediately preceding the recorded BIA decision, immigration judges' grant rates from the decision date through six days afterward rise by 1.57 percentage points more after surprise reversals than after expected reversals. The displayed pre-event coefficients are imprecisely estimated and do not follow a systematic pre-trend. In the preferred unweighted series, the estimate is negative at $k = 1$ and positive at $k = 2$ and in the pooled $k = 3, \ldots, 6$ period, with substantial uncertainty around these later coefficients. The event-time pattern therefore shows the immediate $k = 0$ response more clearly than a persistent post-event path.

The event-equal-weighted estimate at $k = 0$ is also positive and larger than the preferred unweighted estimate. The later weighted estimates remain positive, whereas the preferred unweighted series varies in sign, so the weighting check preserves the direction of the immediate response without implying identical later dynamics. In Appendix Figure A.7, the corresponding unbinned estimates retain a positive differential response at $k = 0$ under both weighting schemes, while estimates farther from the event vary more across periods and are less precise. Appendix Figure A.8 compares the unweighted, unbinned estimates under nearest-event attribution and pre-event-only deduplication. The positive $k = 0$ differential remains robust under both rules. This comparison shows that the immediate response is not specific to the preferred overlap-attribution rule.

Figure 3: Lower-Court Grant-Rate Responses: Surprise versus Expected Reversals



Notes: The figure plots $\hat{b}_\kappa$ from Specification (4.2). Each coefficient measures how the change in lower-court grant rates around surprise reversals differs from the corresponding change around expected reversals, relative to the omitted period $k = -1$. The feedback-diagnosticity event sample contains realized reversals with valid classifier assignments: an expected reversal is one for which the classifier predicts an appeal grant, and a surprise reversal is one for which the classifier predicts an appeal denial. Lower-court cases are linked to events using the nearest-event-attribution rule described in Section 4.1. The x-axis indexes binned seven-day periods around the recorded BIA decision date; period $k = 0$ begins on that date and ends six days later. All specifications include BIA event fixed effects, BIA decision-year-by-day-of-week fixed effects, and BIA decision-month-by-day-of-week fixed effects. Standard errors are clustered by immigration judge. Solid dots with shaded pointwise 95% confidence intervals report the preferred unweighted estimates; open circles with capped pointwise 95% confidence intervals report the event-equal-weighted estimates.

These estimates show that immigration judges adjust their grant rates more strongly after surprise reversals than after expected reversals. Because the two groups share the same formal appellate direction but differ in model-based diagnosticity, this differential adjustment links immigration judges' responsiveness to the diagnostic content of appellate feedback relative to the observable ex ante reversal-risk benchmark. However, the grant-rate adjustment raises an important question: does it extend to observable features of how immigration judges process subsequent cases, or is the response concentrated in the grant decision itself? We next examine early predictability and hearing-process outcomes using the same surprise-versus-expected comparison.

4.3 Observable Hearing-Process Responses

We use two families of observable outcomes as proxies for lower-court hearing diligence. The first is early predictability, which measures how closely final rulings correspond to predictions based on information fixed before the hearing process. The second consists of procedural measures constructed from the administrative hearing records.

Early predictability of lower-court rulings. Section 3.1 constructs five random-forest models that predict lower-court decisions using information fixed before the hearing process. Here, we use specification (5), the most inclusive and preferred specification, to generate the out-of-sample predictions for this analysis. Its predictors are immigration-judge identifier, applicant nationality, case-opening year-month, and filing year-month.

We construct two early-predictability measures from these predictions. The binary correct-prediction measure equals one when the model's prediction matches the observed lower-court ruling. Let $\Pr_{ep,i}(\text{grant})$ denote the predicted grant probability for proceeding i. The continuous early-predictability score is

$$EP_i = 1 - \left| \underset{ep,i}{\Pr}(\text{grant}) - \mathbf{1}\left(\text{Grant}_i = 1\right) \right|.$$

For a granted proceeding, the score equals the predicted grant probability; for a denied proceeding, it equals one minus that probability. The score ranges from zero

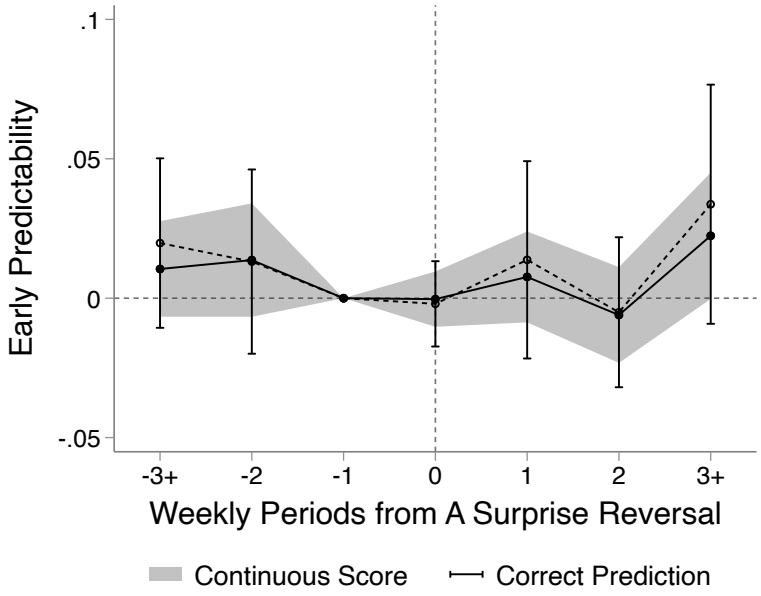
to one, with higher values indicating that the final ruling was more predictable from information fixed before the hearing process.

Using the same feedback-diagnosticity event sample, we re-estimate Specification (4.2) for linked lower-court proceedings with a valid early-predictability measure. Figure 4 presents the preferred unweighted estimates for both measures. Each coefficient $\hat{b}_\kappa$ compares the change in early predictability around surprise reversals with the corresponding change around expected reversals, relative to the omitted period $k = -1$. The binary correct-prediction measure and the continuous early-predictability score produce similar coefficient paths. Neither measure shows a systematic post-event increase after surprise reversals relative to expected reversals. On this diligence proxy, we therefore find no systematic decline in lower-court hearing diligence following surprise reversals. For subsequent analysis, we focus on the continuous score because it preserves variation in the model-predicted probability and provides more precise estimates.

The pre-event estimates fluctuate around zero and are imprecise, so we do not interpret them as evidence that immigration judges anticipate either class of reversal. Appendix Figure A.9 reports the continuous early-predictability score under the general appeal grant-versus-denial specification. Appendix Figure A.10 reports the unbinned surprise-versus-expected estimates under the unweighted and event-equal-weighted specifications. Neither exercise changes the main finding that early predictability shows no systematic post-event increase.

Procedural measures of hearing diligence. The early predictability measures provide outcome-based proxies for lower-court hearing diligence. We complement them with four procedural proxies constructed from the administrative hearing records. These measures capture observable features of how immigration judges conduct and schedule subsequent hearings. For each immigration judge j and date t, we construct the procedural measures described below and link the resulting judge-day observation to eligible BIA event e. The weekly event-time index $k \in \{-6, \ldots, 6\}$ records the position of date t relative to the recorded BIA decision

Figure 4: Lower-Court Early-Predictability Responses: Surprise versus Expected Reversals



Notes: The figure plots $\hat{b}_\kappa$ from Specification (4.2) using the continuous early-predictability score and binary correct-prediction measure defined in Section 4.3. Each coefficient measures how the change in the indicated outcome around surprise reversals differs from the corresponding change around expected reversals, relative to the omitted period $k = -1$. The feedback-diagnosticity event sample contains realized reversals with valid classifier assignments: an expected reversal is one for which the classifier predicts an appeal grant, and a surprise reversal is one for which the classifier predicts an appeal denial. The estimation further requires a valid early-predictability measure for the linked lower-court proceeding. Lower-court proceedings are linked to events using the nearest-event-attribution rule described in Section 4.1. The x-axis indexes binned seven-day periods around the recorded BIA decision date; period $k = 0$ begins on that date and ends six days later. All specifications are unweighted and include BIA event fixed effects, BIA decision-year-by-day-of-week fixed effects, and BIA decision-month-by-day-of-week fixed effects. Standard errors are clustered by immigration judge. Solid dots with shaded pointwise 95% confidence intervals report estimates for the continuous score; open circles with capped pointwise 95% confidence intervals report estimates for the binary correct-prediction measure.

date. Using the binned event-time set K defined above, we estimate

$$Y_{jtek} = \sum_{\kappa \in K} \left[\theta_\kappa \mathbf{1}(k \in \kappa) + \delta_\kappa \mathbf{1}(k \in \kappa) \cdot \mathbf{1}\left(\text{Surprise Reversal}_e = 1\right)\right] + \mathbf{X}_e' \boldsymbol{\varphi} + \varepsilon_{jtek}, \tag{4.3}$$

where, Y_{jtek} is one of the four judge-day procedural measures for immigration judge j on date t, linked to BIA event e at event time k. Period $k = -1$ is omitted. The coefficients θ_κ trace the outcome path around expected reversals, while δ_κ measures how the corresponding path around surprise reversals differs from that

path relative to the omitted period. The vector $\mathbf{X}_e$ contains the same BIA event and calendar fixed effects as Specification (4.2), and standard errors are clustered by immigration judge.

For each proceeding associated with a case in the main analysis, we construct four procedural measures from the hearing records. The first is the number of hearing records for that proceeding. The second is average adjournment duration, measured in minutes, across records with valid start and stop times. The third is the share of records with a nonmissing adjournment-reason classification that are coded as preparation adjournments. The fourth is the share of records with a nonmissing calendar-type classification that are coded as individual-calendar hearings.[19] Appendix Table A.6 reports proceeding-level summary statistics for the four procedural measures from 1984 through 2019. Panel A includes proceedings associated with cases in the main analysis, while Panel B includes all proceedings in the hearing records. In Panel A, proceedings average 3.79 hearing records, an adjournment duration of 139 minutes, a preparation-adjournment share of 8%, and an individual-calendar share of 35%.[20] For the event-study analysis, we first construct each measure within proceeding-by-judge-by-date cells and then average across the eligible proceedings handled by the same immigration judge on that date. This construction gives each proceeding equal weight within the resulting judge-day measure. We link these judge-day measures to feedback-diagnosticity events and apply the nearest-event-attribution rule described in Section 4.1. The regression unit is therefore a judge-day-event-period observation. Because valid clock times, adjournment-reason classifications, and calendar-type classifications are not available for every hearing record, the estimation sample varies across the four outcomes.

Table 1 presents the preferred unweighted estimates. In Column (1), the $k = 0$ differential is 0.017 hearing records per proceeding. Thus, from the recorded BIA decision date through the following six days, hearing records per proceeding rise by 0.017 more after surprise reversals than after expected reversals, relative to pe-

[19] A preparation adjournment records additional time granted for case preparation. Individual-calendar hearings address the evidence and issues in a particular proceeding, whereas master-calendar hearings generally focus on pleadings, scheduling, and other preliminary matters. The two shares therefore capture how much of the recorded hearing process is allocated to preparation and individualized adjudication, making them useful procedural proxies for hearing diligence.

[20] We winsorize the recorded adjournment-duration measure at the upper 1% of the full hearing-record distribution before constructing the proceeding-level averages.

riod $k = -1$. The estimate is significant at the 10% level and is consistent with a modest increase in hearing diligence on this procedural margin. The later point estimates remain positive but imprecise, so the evidence is strongest for an immediate response. Appendix Figure A.11 shows that the $k = 0$ estimate remains positive under event-equal weighting. Column (2) shows that the surprise-versus-expected difference in adjournment duration is already present before the event. The differential is 6.126 minutes in the early pre-event bin and 3.731 minutes at $k = -2$, and both estimates are statistically distinguishable from zero. At $k = 0$, the differential is minor and imprecise. Among the later differentials, only the $k = 2$ coefficient is statistically distinguishable from zero. Because these differences are already present before the event, the later positive coefficients do not identify a post-reversal response. Columns (3) and (4) show that, at $k = 0$, the preparation-adjournment-share differential is -0.009 and the individual-calendar-share differential is 0.013. Both are imprecisely estimated. These procedural proxies therefore show no systematic adjustments in hearing diligence on either compositional margin after surprise reversals relative to expected reversals.

The hearing records also cover non-asylum proceedings, with records outside the cases associated with the main analysis accounting for roughly 81% of the broader hearing dataset. To examine whether the response extends to judges' broader hearing behavior, we reconstruct the four judge-day procedural measures using all observed proceedings and re-estimate Specification (4.3). Appendix Table A.7 reports that hearing records per proceeding, adjournment duration, and the preparation-adjournment share show no systematic differences between surprise and expected reversals. The individual-calendar-share estimates are positive throughout the post-event periods, although the immediate estimate of 0.012 is imprecise. The pooled late post-event estimate is 0.016 and significant at the 10% level. The broader-sample evidence therefore points to a modest shift toward individual-calendar hearings rather than a broad change across procedural measures.

Taken together, these results show that the stronger grant-rate response to surprise reversals is not accompanied by a broad decline in our proxies for lower-court hearing diligence. Early predictability shows no systematic post-event increase, while the number of hearing records per proceeding rises modestly in the first post-event week. The other procedural measures do not show a systematic im-

Table 1: Lower-Court Procedural Responses: Surprise versus Expected Reversals

	(1) No. hearings	(2) Avg. duration	(3) Prep. share	(4) Ind. cal. 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)
$\mathbf{1}(\text{Surprise Reversal}_e = 1) \times \mathbf{1}(k \in [-6,-3])$	0.010	6.126***	-0.007	-0.011
	(0.010)	(1.972)	(0.011)	(0.014)
$\mathbf{1}(\text{Surprise Reversal}_e = 1) \times \mathbf{1}(k = -2)$	0.010	3.731*	-0.008	0.014
	(0.013)	(2.003)	(0.009)	(0.011)
$\mathbf{1}(\text{Surprise Reversal}_e = 1) \times \mathbf{1}(k = 0)$	0.017*	1.881	-0.009	0.013
	(0.010)	(1.564)	(0.008)	(0.010)
$\mathbf{1}(\text{Surprise Reversal}_e = 1) \times \mathbf{1}(k = 1)$	0.005	1.386	0.001	0.013
	(0.010)	(1.671)	(0.008)	(0.010)
$\mathbf{1}(\text{Surprise Reversal}_e = 1) \times \mathbf{1}(k = 2)$	0.018	4.475**	-0.007	0.007
	(0.012)	(2.232)	(0.010)	(0.010)
$\mathbf{1}(\text{Surprise Reversal}_e = 1) \times \mathbf{1}(k \in [3,6])$	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

Notes: This table reports unweighted estimates of $\hat{\theta}_\kappa$ and $\hat{\delta}_\kappa$ from Specification (4.3) for the four judge-day procedural measures defined in Section 4.3. The $\hat{\theta}_\kappa$ rows trace changes around expected reversals relative to the omitted period $k = -1$. The $\hat{\delta}_\kappa$ rows measure how the corresponding changes around surprise reversals differ from those around expected reversals, relative to the same omitted period. The estimation sample links judge-day measures constructed from proceedings associated with cases in the main analysis to feedback-diagnosticity events from 1986 through 2019. Lower-court judge-day observations are linked to events using nearest-event attribution. All specifications include BIA event fixed effects, BIA decision-year-by-day-of-week fixed effects, and BIA decision-month-by-day-of-week fixed effects. Standard errors clustered by immigration judge are reported in parentheses. The number of judge-day-event-period observations differs across columns because the underlying hearing fields have different rates of valid reporting. Significance level: 1% (***), 5% (**), 10% (*).

mediate change. These process measures do not distinguish belief updating from strategic responses intended to reduce future reversals. The grant-rate and process evidence therefore points to a substantive adjustment in lower-court decision-

making after more diagnostic appellate feedback, without a broad retreat from hearing diligence along the observed margins.

The event studies show that immigration judges respond more strongly to more diagnostic appellate feedback. The deeper question is whether the judges who respond most strongly are also more closely aligned with the BIA standard. Section 5 develops a framework linking response magnitude to alignment and compares high-response and low-response judges across the distribution of predicted grant probability.

5 Attention, Learning, and Distributional Alignment

This section examines how attention and learning shape immigration judges' responses to appellate feedback and whether stronger responses are associated with closer distributional alignment with the BIA standard. We develop a threshold model showing how reversal rates depend on both the location of a judge's decision threshold and the strength of the judge's response to feedback. In the empirical analysis, we use judge-specific response magnitude as a proxy for the model's recipient-side attentiveness and effective updating, then use marginal treatment effects (MTEs) to compare high-response and low-response judges across the distribution of predicted lower-court grant probability.

5.1 A Simple Model of Reversals

Each case i has a latent quality $q_i \in [0, 1]$ that summarizes its strength under the BIA standard. The distribution of q_i has density $f(q) > 0$ and CDF $F(q)$, with f and F continuously differentiable. An immigration judge j with leniency level l_j applies a threshold $c_j = c(l_j)$, where $c(\cdot)$ is differentiable and $c'(l) < 0$. Thus, more lenient judges apply lower grant thresholds. Judge j grants asylum if and only if

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

An applicant whose case is denied may file an appeal. We represent appeal filing with the cutoff rule

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

where $a \in [0, c^*)$ is the appeal-filing threshold.[21] Among lower-court denials, applicants with sufficiently high case quality therefore find it worthwhile to file an appeal. This representation nests richer models in which the benefit of appeal is increasing in q_i. The BIA applies a fixed standard $c^* \in (0, 1)$ and grants an appeal if and only if $q_i \geq c^*$. We focus on the empirically relevant region in which immigration judges apply stricter thresholds than the BIA, so $c_j > c^*$.

A reversal occurs when an immigration judge denies a case that the BIA grants on appeal. Let

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

denote a reversal in case i. The model-implied reversal rate among all cases assigned to judge j with cutoff c_j is

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

Thus, $R(c_j)$ is the unconditional reversal probability among cases assigned to judge j.

The unconditional reversal rate factors into the probability of a lower-court denial and two conditional review margins:

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

where $D_i = 0$ denotes a lower-court denial, $A_i = 1$ denotes an appeal following that denial, and $S_i = 1$ denotes a BIA grant among appealed denials. We call the appeal rate among lower-court denials, $E(c_j) \coloneqq \Pr(A_i = 1 \mid D_i = 0, c_j)$, the appeal-filing, or selection (*extensive*), margin; and the appeal grant rate among appealed denials, $I(c_j) \coloneqq \Pr(S_i = 1 \mid A_i = 1, D_i = 0, c_j)$, the BIA-success (*intensive*) margin.

Under the cutoff rules,

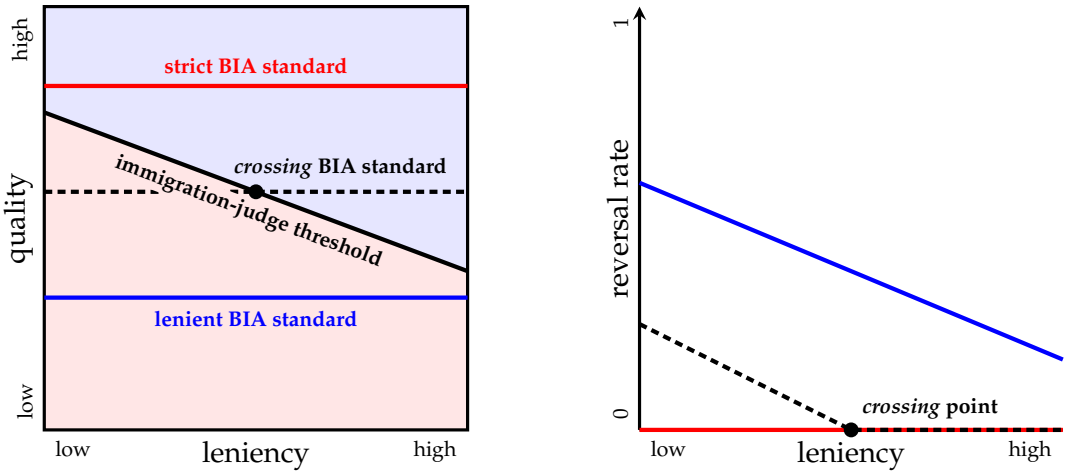
$$E(c_j) = \frac{F(c_j) - F(a)}{F(c_j)}, \qquad I(c_j) = \frac{F(c_j) - F(c^*)}{F(c_j) - F(a)}.$$

[21] Equivalently, the cost of appeal may decrease with q_i. The cutoff rule provides a simple representation of this relationship.

The corresponding reversal rate conditional on a lower-court denial is

$$R^D(c_j) \coloneqq \Pr(Y_i = 1 \mid D_i = 0, c_j) = \frac{F(c_j) - F(c^*)}{F(c_j)} = E(c_j)I(c_j).$$

Figure 5: A Stylized Illustration of Judge Leniency and Reversals



(a) Decision Thresholds

(b) Unconditional Reversals

Notes: Panel 5a shows how the immigration-judge threshold declines with leniency under three possible positions of the BIA standard. Panel 5b shows the corresponding unconditional reversal rate $R(c_j)$. Coordinates and slopes are schematic and chosen for expositional clarity.

Figure 5 provides a stylized illustration of the model's central intuition: reversals depend on the position of immigration-judge thresholds relative to the BIA standard. Panel 5a shows the decision thresholds, and Panel 5b traces the corresponding unconditional reversal rates. When the BIA standard is stricter than every immigration-judge threshold, no lower-court denial satisfies the BIA grant standard, so the reversal rate is zero. When the BIA standard is more lenient, stricter immigration judges deny more cases that satisfy the BIA standard, and the reversal rate falls as judge leniency increases. In the crossing case, the same relationship holds until the immigration-judge threshold reaches the BIA standard, after which the reversal rate remains at zero. Proposition 5.1 formalizes this comparative static in the region $c_j > c^*$ and extends it to the denial-conditioned reversal rate and both review margins:

Proposition 5.1. *For judges with $c_j > c^*$ and appeal-filing threshold $a < c^*$, the unconditional reversal rate $R(c_j)$, the denial-conditioned reversal rate $R^D(c_j)$, and the success*

margin $I(c_j)$ are strictly decreasing in l_j. The selection margin $E(c_j)$ is weakly decreasing in l_j and strictly decreasing whenever $a > 0$.

The intuition follows from the cases between the BIA standard and the immigration judge's threshold. A stricter threshold expands the denial pool to include more cases that satisfy the BIA standard, increasing both reversal rates and the BIA grant rate among appealed denials. When $a > 0$, it also raises the share of denials that are appealed. When $a = 0$, every denial satisfies the appeal-filing cutoff, so the selection margin remains constant. Greater leniency reverses these relationships by lowering the threshold and removing these cases from the denial pool. Detailed proofs are provided in Section C.

The static model treats each judge's threshold as fixed. We now allow appellate feedback conveyed by reversals to change the threshold that an immigration judge applies in future cases. Let $c_{j,t}$ denote judge j's threshold in period t. We represent this updating process as

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

Here, $\lambda_j \in [0, 1]$ is a judge-specific effective updating rate, with attention as one of its components. It absorbs whether a judge notices and understands the appellate signal, regards it as relevant, retains it, and is willing and institutionally able to apply it to future decisions. It can also reflect organizational frictions in the transmission and implementation of BIA feedback. A judge with $\lambda_j = 0$ leaves the threshold unchanged, while $\lambda_j > 0$ moves it toward c^*, with larger values producing faster adjustment. We use *inattentive* and *attentive* as theoretical shorthand for zero and positive effective updating, respectively. The prediction model characterizes the signal side of learning by classifying the diagnosticity of appellate feedback, while λ_j characterizes the recipient side by governing the strength of the behavioral response.

Iterating the updating law gives

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

This threshold path yields a straightforward learning result:

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

standard:

$$\lim_{t\to\infty} c_{j,t} = c^*.$$

For an inattentive judge with $\lambda_j = 0$, the threshold remains fixed at its initial value:

$$c_{j,t} = c_{j,0} > c^* \quad \textit{for all } t.$$

The proposition formalizes the idea that positive effective updating closes the gap between the immigration judge's threshold and the BIA standard, while zero effective updating preserves the initial gap. This convergence describes alignment with the modeled BIA standard c^*.

To connect threshold updating to reversal outcomes, we examine how the unconditional reversal rate, the denial-conditioned reversal rate, and both review margins respond locally to a change in the decision threshold. Define

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

Each function measures how strongly the corresponding outcome responds to a marginal increase in threshold strictness. The following shape restrictions compare these sensitivities at the initial strict threshold $c_{j,0}$ and the BIA-aligned threshold c^*:

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

$$\psi_R(c_{j,0}) > \psi_R(c^*), \qquad \psi_{R^D}(c_{j,0}) > \psi_{R^D}(c^*), \qquad \psi_E(c_{j,0}) \geq \psi_E(c^*),$$

where the final inequality is strict whenever $a > 0$. If, in addition, $f'(c)\left[F(c) - F(a)\right] > 2f(c)^2$ for all $c \in [c^*, c_{j,0}]$, then

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

The sensitivity ranking in Lemma 5.1 follows from the shape of the case-quality distribution over $[c^*, c_{j,0}]$. The density at the threshold rises sufficiently quickly relative to the cumulative mass of denials and appealable denials that a marginal change in a stricter threshold reclassifies more relevant cases. Under the first set of

conditions, the unconditional and denial-conditioned reversal rates respond more strongly at $c_{j,0}$ than at c^*. The same ranking holds for the selection margin when $a > 0$; when $a = 0$, every denial is appealed and its sensitivity is zero at both thresholds. The additional condition delivers the corresponding ranking for the BIA-success margin. Positive effective updating moves an attentive judge's threshold toward the lower-sensitivity region around the BIA standard, while an inattentive judge remains at the higher-sensitivity initial threshold. The model therefore predicts flatter relationships between leniency and reversal outcomes for judges whose thresholds lie closer to the BIA standard. Section C proves Lemma 5.1 and discusses the required shape restrictions.

The model highlights two dimensions of judge heterogeneity that organize the empirical analysis: attentiveness to appellate feedback and leniency in lower-court decisions. We first estimate judge-specific behavioral responses to surprise rather than expected reversals and use response magnitude as a proxy for recipient-side attention and effective updating. We then construct a leave-out measure of judge leniency and compare how reversal outcomes vary with leniency for high-response and low-response judges. Together, these measures allow us to assess whether stronger responsiveness is associated with closer distributional alignment with the BIA standard.

5.2 Measuring Responsiveness and Leniency

The model establishes how effective updating and leniency shape reversal outcomes and distributional alignment with the BIA standard. We now construct the two measures used in the empirical analysis: behavioral responsiveness to appellate feedback, based on judge-specific changes in lower-court behavior after surprise relative to expected reversals, and leave-out judge leniency, which accounts for the case-assignment process.

5.2.1 Behavioral Responsiveness to Appellate Feedback

For each immigration judge j, we re-estimate Specification (4.2) using the collapsed event-time bins $K' = \{[-6,-3],\{0,1\},[2,6]\}$:

$$\begin{aligned} Y_{ijek} =& \eta_{1j}\mathbf{1}(k \in [-6,-3]) + \mu_{1j}\mathbf{1}(k \in [-6,-3])\mathbf{1}\left(\text{Surprise Reversal}_e = 1\right) \\ &+ \eta_{2j}\mathbf{1}(k \in \{0,1\}) + \mu_{2j}\mathbf{1}(k \in \{0,1\})\mathbf{1}\left(\text{Surprise Reversal}_e = 1\right) \\ &+ \eta_{3j}\mathbf{1}(k \in [2,6]) + \mu_{3j}\mathbf{1}(k \in [2,6])\mathbf{1}\left(\text{Surprise Reversal}_e = 1\right) \\ &+ \mathbf{X}_e'\boldsymbol{\psi}_j + \nu_{ijek}. \end{aligned} \tag{5.1}$$

The surprise-reversal indicator equals one for a realized reversal that the classifier predicted would be an appeal denial and zero for a realized reversal that it predicted would be an appeal grant. The coefficients μ_{mj} therefore measure how judge j's outcome path around surprise reversals differs from the corresponding path around expected reversals, relative to the omitted pre-event category. The vector $\mathbf{X}_e$ contains the same BIA event and calendar fixed effects as Specification (4.2). Periods $k = -2$ and $k = -1$ form the omitted category. We use $\hat{\mu}_{2j}$ as judge j's behavioral-response coefficient. It measures the differential change in the outcome after surprise rather than expected reversals during the two-week window beginning on the recorded BIA decision date, relative to the omitted pre-event periods.[22]

Grant-rate responsiveness. For this measure, Y_{ijek} is an indicator equal to one if immigration judge j grants proceeding i. Appendix Figure A.12 plots $\hat{\mu}_{2j}$ for the 354 immigration judges with estimable grant-rate responses. The estimates vary substantially across judges and include both positive and negative responses, although their mean and median are positive, consistent with the aggregate event-study result. The sign records the direction of the average grant-rate change, while the absolute value isolates the strength of the response without assigning that direction to an unobserved threshold bias. We therefore use $|\hat{\mu}_{2j}|$ as our primary measure of grant-rate responsiveness. In the language of the model, response magnitude proxies variation in the composite recipient-side effective updating repre-

[22]The pooled window $k \in \{0,1\}$ covers the recorded BIA decision date through thirteen days afterward. Pooling these periods improves precision in the judge-specific estimates while retaining separate controls for earlier leads and later lags.

sented by λ_j rather than directly estimating λ_j.

Early-predictability responsiveness. We construct two secondary measures of how immigration judges adjust their hearing processes after surprise relative to expected reversals. The first uses early predictability as an outcome-based proxy for hearing diligence. For this measure, Y_{ijek} is the continuous early-predictability score defined in Section 4.3. We re-estimate Specification (5.1) separately for each judge and use the raw coefficient $\hat{\mu}_{2j}$ as the judge's early-predictability response. A negative coefficient indicates that the judge's final rulings become less predictable from information fixed before the hearing process during the two-week post-event window. This pattern is consistent with a greater role for information developed during the hearing process. Appendix Figure A.13 plots the responses for 380 judges with estimable coefficients.

Hearing-count responsiveness. Our second measure of hearing-process adjustment captures changes in the average number of hearings per proceeding. We apply the same collapsed-bin design to Specification (4.3) and estimate the following regression separately for each immigration judge j:

$$\begin{aligned} Y_{jtek} =& \eta_{1j}\mathbf{1}(k \in [-6,-3]) + \mu_{1j}\mathbf{1}(k \in [-6,-3])\mathbf{1}\left(\text{Surprise Reversal}_e = 1\right) \\ &+ \eta_{2j}\mathbf{1}(k \in \{0,1\}) + \mu_{2j}\mathbf{1}(k \in \{0,1\})\mathbf{1}\left(\text{Surprise Reversal}_e = 1\right) \\ &+ \eta_{3j}\mathbf{1}(k \in [2,6]) + \mu_{3j}\mathbf{1}(k \in [2,6])\mathbf{1}\left(\text{Surprise Reversal}_e = 1\right) \\ &+ \mathbf{X}_e'\boldsymbol{\psi}_j + \nu_{jtek}. \end{aligned} \tag{5.2}$$

Here, Y_{jtek} is judge j's daily average number of hearing records per proceeding on date t, linked to BIA event e in event-time period k. Periods $k = -2$ and $k = -1$ remain the omitted category. We use the raw coefficient $\hat{\mu}_{2j}$ as judge j's hearing-count response. A positive coefficient indicates a larger increase in hearing records per proceeding after surprise rather than expected reversals during the two-week post-event window. We interpret this increase as a stronger response in the procedural proxy for hearing diligence. Appendix Figure A.14 plots the responses for 416 judges with estimable coefficients.

Appendix Figure A.15 plots the pairwise relationships among the three judge-level responsiveness measures. The weak correlations indicate that the three re-

sponsiveness measures capture empirically distinct dimensions of behavioral adjustment to appellate feedback. Appendix Figure A.16 shows a similar pattern when the measures are constructed at the judge-year level.

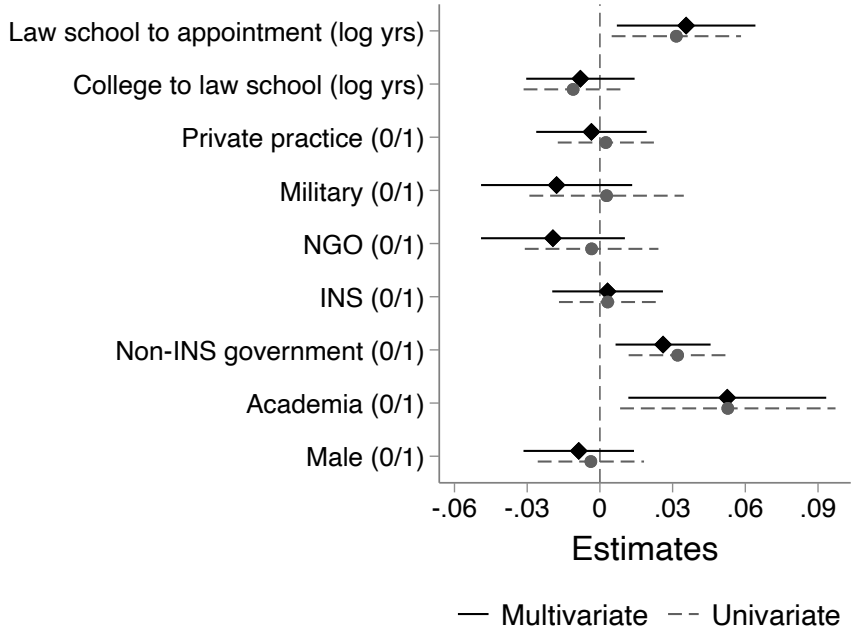
Patterns of judicial attentiveness. We use the three responsiveness measures as proxies for immigration judges' attentiveness to appellate feedback. Grant-rate response magnitude is our primary proxy because it captures how strongly judges change their ruling behavior after surprise rather than expected reversals. Through the model, we interpret stronger responsiveness as greater lower-court judge attention and effective updating.

For the main alignment analysis, we rank proceedings within each lower-court ruling year by the grant-rate response magnitude of the presiding judge. Proceedings in the upper tercile are assigned to high-response judges, proceedings in the lower tercile are assigned to low-response judges, and the middle tercile is omitted from the main comparison. For the hearing process measures, we interpret a decline in early predictability and an increase in hearing counts as greater attentiveness along the corresponding dimension.

We next relate responsiveness to immigration judges' biographical characteristics. At the judge level, Figure 6 shows that grant-rate responsiveness is positively associated with years of professional experience between law school graduation and appointment and with prior non-INS government or academic experience. Among the secondary measures, prior NGO experience is associated with a larger decline in early predictability, while prior military and NGO experience are associated with lower hearing-count responses in the multivariate specification (Appendix Figure A.17).[23] These associations yield no consistent ranking across measures or aggregation levels. The three judge-year responsiveness measures also remain relatively stable across presidential periods, with no statistically significant differences across administrations (Appendix Figures A.20, A.21, and A.22).

[23]The judge-year estimates identify additional relationships with professional background and gender, although these patterns differ from the judge-level estimates (Appendix Figures A.18 and A.19).

Figure 6: Grant-Rate Responsiveness and Judge Characteristics



Notes: The figure reports associations between the judge-level grant-rate responsiveness measure and immigration 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.

5.2.2 Judge Leniency

Following Dobbie, Goldin and Yang (2018) and Arnold, Dobbie and Yang (2018), we construct a residualized leave-out judge-leniency measure. Judge leniency provides the empirical counterpart to the location of an immigration judge's decision threshold. In the model, greater leniency corresponds to a lower decision threshold and a greater tendency to grant asylum. The constructed judge-leniency measure provides the source of variation for the subsequent analysis.

Specifically, we residualize the lower-court grant indicator using court-by-time fixed effects that account for the case-assignment process.[24] For each proceeding, we then calculate the presiding immigration judge's mean residual grant rate among other cases decided by that judge in the same year, excluding all proceed-

[24]The specification absorbs court-by-decision-year-by-day-of-week and court-by-month-of-year-by-day-of-week fixed effects, the levels at which cases are quasi-randomly assigned to judges.

ings associated with the same case. Higher values indicate a greater tendency to grant after accounting for the case-assignment controls.

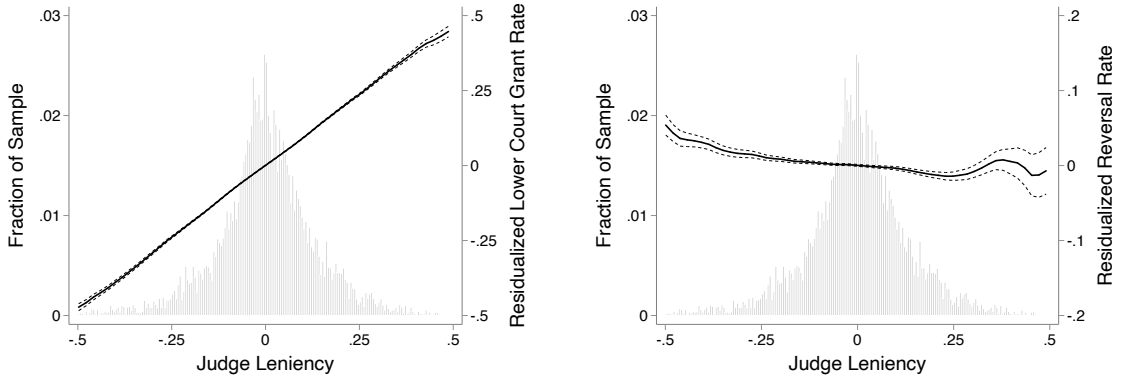
Existence of first stage. We first establish that the leave-out judge-leniency measure strongly predicts lower-court grant decisions. We residualize the lower-court grant indicator using the same court-by-time fixed effects and plot its local linear relationship with judge leniency. Figure 7a shows that the residualized grant rate increases steadily with judge leniency, providing the first-stage variation used in the subsequent analysis.

Exclusive restriction. Next, we test whether observed proceeding characteristics remain associated with judge leniency after conditioning on the court-by-time assignment cells. Appendix Table A.8 reports the results of this randomization test. Column (1) shows that applicant, case, assignment-timing, and decision-process characteristics jointly predict lower-court grants. Column (2) shows that the characteristics are jointly significant, but most administrative complexity and assignment timing measures are not significantly related to leniency. The remaining associations are concentrated in applicant and proceeding characteristics, including representation, procedural type, and origin region. As in earlier judge-leniency studies of racial bias (Arnold, Dobbie and Yang, 2018; Dobbie, Goldin and Yang, 2018), the correlations with applicant origin and procedural type may reflect substantive differences in how judges evaluate cases across groups. They motivate our heterogeneity analyses by applicant origin and procedural setting, which maintain the same identifying conditions within each subgroup.

Monotonicity. The final assumption we check is monotonicity: assignment to a more lenient immigration judge should not turn a grant into a denial. Equivalently, a case granted by a stricter judge would also be granted by a more lenient judge, while a case denied by a more lenient judge would also be denied by a stricter judge. This condition is required for the local instrumental variables interpretation (LIV) of the MTEs (Angrist, Imbens and Rubin, 1996; Heckman and Vytlacil, 2005). A useful empirical diagnostic is rank stability: judges who are relatively lenient for one observable group should also be relatively lenient for the corresponding comparison group. In Appendix Figure A.23, we reconstruct the leave-out le-

niency measure separately across applicant-origin and case-characteristic groups. The correlations between the group-specific measures are positive and statistically significant in every comparison, providing rank-stability evidence consistent with monotonicity.

Figure 7: Judge Leniency, Lower-Court Grants and Reversals



(a) Lower-Court Grant Rate

(b) Reversal Rate among Denials

Notes: The figure relates the leave-out judge-leniency measure to lower-court grant and reversal outcomes. Panel 7a shows a local linear fit of the residualized lower-court grant indicator on judge leniency. Panel 7b shows the corresponding relationship for the reversal indicator among lower-court denials. The leave-out judge-leniency measure is constructed as described in Section 5.2.2. The reversal indicator equals one when a lower-court denial receives a BIA grant and zero for every other lower-court denial, including denials that are not appealed. Both outcomes are residualized using the same court-by-time fixed effects used to construct the leniency measure. The gray bars show the distribution of judge leniency across proceedings, and the dashed lines denote pointwise 95% confidence intervals.

5.3 Testing Model Predictions

Having constructed measures of judge responsiveness and leniency, we first study the reduced-form relationships between leniency and the denial-conditioned reversal rate, appeal-filing margin, and BIA-success margin. We then use marginal treatment effects (MTEs) to determine whether high-response judges exhibit closer distributional alignment with the BIA standard than low-response judges.

5.3.1 Reduced-Form Evidence

The model in Section 5.1 delivers clear predictions. Proposition 5.1 predicts that the denial-conditioned reversal rate $R^D(c)$ and the BIA-success margin decrease with judge leniency when immigration judges apply thresholds stricter than the

BIA standard. The selection margin is weakly decreasing and decreases strictly when $a > 0$. Lemma 5.1 provides conditions under which these declining relationships are steeper in the strict region than near the BIA-aligned threshold. Figure 7b shows that the residualized reversal rate declines with judge leniency and becomes flatter at higher levels of leniency. This is consistent with the model: stricter judges retain more cases that satisfy the BIA standard in their denial pools, while greater leniency removes these cases through first-instance grants.

We then examine the two review margins through which BIA feedback is generated. Panel A.24a of Appendix Figure A.24 plots the appeal-filing margin, defined among all lower-court denials, against judge leniency. Panel A.24b plots the BIA-success margin, defined among appealed denials with a classified BIA grant or denial disposition. Both relationships are negative: denials issued by more lenient judges are less likely to be appealed, and appealed denials issued by more lenient judges are less likely to receive a BIA grant. These distinct conditional relationships show that judge leniency is associated with both entry into the appellate-review channel and the BIA disposition once an appeal reaches the BIA.

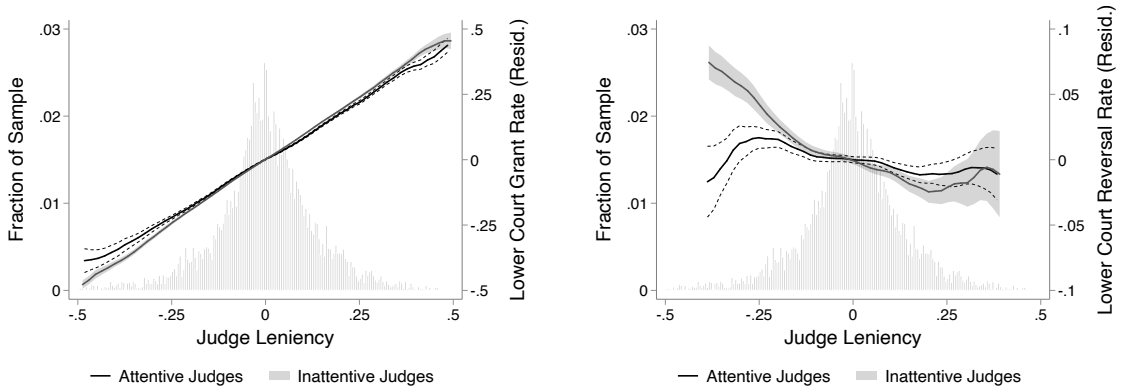
We next examine whether observable BIA grant risk varies with judge leniency. Using the random-forest model described in Section 3.2, we assign each lower-court proceeding a predicted BIA grant probability based on information available before appellate resolution. Appendix Figure A.25 plots the residualized predicted probability against judge leniency and shows a modest positive relationship: proceedings decided by more lenient judges have slightly higher predicted BIA grant probabilities. We interpret this relationship as a case-composition diagnostic, indicating that observable characteristics associated with BIA grants vary modestly with judge leniency. This modest relationship supports treating the BIA standard c^* as independent of immigration-judge leniency in the stylized model.[25]

Proposition 5.2 and Lemma 5.1 predict that stronger effective updating moves immigration-judge thresholds toward the BIA standard, where reversal outcomes are less sensitive to leniency. Using the grant-rate response-magnitude classification defined in Section 5.2.1, we refer to high-response judges as *attentive* and low-response judges as *inattentive*, linking our primary empirical proxy for attentiveness to the model.

[25] Appendix Figure A.26 incorporates this modest relationship into the stylized model and shows that it flattens, but does not reverse, the negative relationship between judge leniency and reversals.

Grant-rate attentiveness. Figure 8 compares the reduced-form relationships for proceedings ruled by attentive and inattentive judges. In Panel 8a, residualized lower-court grant rates rise with judge leniency for both groups, showing first-stage variation within each subsample. Panel 8b presents the central comparison. The denial-conditioned reversal rate declines sharply with leniency among inattentive judges but exhibits a substantially flatter relationship among attentive judges. This contrast links greater attentiveness to weaker sensitivity of reversals to individual judge leniency, consistent with closer distributional alignment with the BIA standard. Appendix Figure A.27 shows the corresponding results using judge-year attentiveness measures, and Appendix Figure A.28 uses a top-versus-bottom-half classification. Both exercises preserve the central contrast between attentive and inattentive judges.

Figure 8: Judge Leniency and Lower-Court Outcomes by Grant-Rate Attentiveness



(a) Lower-Court Grant Rate (b) Reversal Rate among Denials

Notes: The figure relates judge leniency to lower-court outcomes separately for attentive and inattentive judges. Attentiveness is measured by the magnitude of each judge's grant-rate response to surprise reversals, as described in Section 5.2.1. Within each ruling year, proceedings in the top tercile of grant-rate response magnitude are classified as attentive, while those in the bottom tercile are classified as inattentive. Panel 8a shows local-linear fits of the residualized lower-court grant indicator on judge leniency. Panel 8b shows the corresponding fits for the reversal indicator among lower-court denials. Both outcomes are residualized using the same court-by-time fixed effects used to construct the leave-out judge-leniency measure described in Section 5.2.2. The gray bars show the distribution of judge leniency across proceedings. The black line represents attentive judges, and the gray line represents inattentive judges; dashed lines and shaded areas denote pointwise 95% confidence intervals.

Hearing-diligence attentiveness. Next, we use judges' early-predictability and hearing-count responses as proxies for attentiveness in the hearing process. A decline in early predictability and an increase in hearings per proceeding indicate

greater attentiveness along the corresponding dimension.

Appendix Figure A.29 compares judges whose early predictability decreases after surprise reversals with those whose early predictability increases. The estimated grant-rate relationships are similar across the two groups. The denial-conditioned reversal rate exhibits a flatter relationship with leniency among judges whose early predictability decreases, but the contrast is less pronounced than for grant-rate attentiveness.

Appendix Figure A.30 makes the corresponding comparison based on whether judges increase or decrease the number of hearings per proceeding after surprise reversals. Both the grant-rate and reversal-rate relationships are similar across these groups.[26] Overall, these secondary measures provide weaker reduced-form separation between attentive and inattentive judges than grant-rate response magnitude. The early-predictability comparison points in the same direction as the primary result, while the hearing-count comparison shows little difference. Taken together, the reduced-form results point to grant-rate responsiveness as the clearest empirical link between attentiveness to appellate feedback and distributional alignment with the BIA standard.

5.3.2 Distributional Alignment: Marginal Treatment Effects

The reduced-form relationship between reversal risk and leniency differs between attentive and inattentive judges but does not show where this difference emerges along the predicted grant-probability distribution. We therefore estimate marginal treatment effects (MTEs) using the local instrumental variables (LIV) framework of Heckman and Vytlacil (2005) and the empirical implementation in Arnold, Dobbie and Yang (2018). The MTEs trace local changes in reversal outcomes across predicted grant probability, allowing us to examine whether stronger responsiveness to appellate feedback is associated with closer distributional alignment with the BIA standard.

Let D_{ijt} denote the lower-court grant decision for case i ruled by judge j at time t, let Z_{ijt} denote the constructed leniency instrument, and let $\mathbf{X}_{ijt}$ collect the court-by-time fixed effects. The first-stage relationship defines the predicted lower-court

[26] The unpublished appendix reports corresponding specifications using judge-year response measures and alternative classifications based on the top and bottom terciles.

grant probability

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

p_{ijt} indexes the component of the lower-court grant probability predicted by judge leniency after accounting for court-by-time differences. Section 5.2.2 documents the corresponding first-stage relationship.

To match the denial-conditioned reversal rate $R^D(c)$ in the model, the main MTE sample consists of lower-court denials. Within this sample, $Y_{ijt} = 1$ if a denial is followed by an appeal grant and $Y_{ijt} = 0$ otherwise. Let $\tilde{Y}_{ijt}$ denote this outcome residualized using the full set of court-by-time fixed effects. We define the MTE as

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

$\text{MTE}(p)$ measures the local change in the expected denial-conditioned reversal outcome as p increases. Because p is constructed as a positive monotone transformation of judge leniency, the sign and shape of $\text{MTE}(p)$ can be compared directly with the model's comparative statics in Section 5.1.

We estimate the MTEs separately for attentive and inattentive judges in three steps. First, following Heckman, Urzua and Vytlacil (2006), we partial out the court-by-time fixed effects used to construct the leniency measure. Within each group, we regress the residualized lower-court grant decision on judge leniency and construct the predicted grant probability $\tilde{p}_{ijt} = \bar{D}_{ijt} + \hat{\theta} Z_{ijt}$.[27] Second, we estimate $E[\tilde{Y}_{ijt} \mid p_{ijt} = p]$ using local linear regression with a bandwidth of 0.08 and obtain its derivative numerically. Third, we repeat the group-specific estimation in 500 judge-cluster bootstrap samples. The pointwise 90% confidence intervals are the 5th and 95th percentiles of the bootstrap estimates at each grid point.

To summarize the estimated functions and facilitate comparisons across judge groups, we complement the pointwise MTE estimates with uniformly weighted averages over a shared trimmed region of the mean-centered predicted grant-

[27] We use the 10th and 90th percentiles of the full-sample predicted grant-probability distribution to define a fixed grid of 200 evaluation points shared across groups and bootstrap replications. Before constructing the evaluation grid, predicted probabilities are bounded between 0.01 and 0.99.

probability grid. Let $\mathcal{P} = [p_L, p_U]$ denote this region. For judge group g, we define

$$\overline{\mathrm{MTE}}_g(\mathcal{P}) \equiv \frac{1}{p_U - p_L} \int_{p_L}^{p_U} \mathrm{MTE}_g(p)\, \mathrm{d}p. \tag{5.5}$$

This estimand assigns equal weight to equal-length intervals in $\mathcal{P}$.[28] For a region $\mathcal{Q} \in \{\mathcal{P}, \mathcal{P}^S\}$, we compare judge groups g and g' using

$$\Delta(\mathcal{Q}) \equiv \overline{\mathrm{MTE}}_g(\mathcal{Q}) - \overline{\mathrm{MTE}}_{g'}(\mathcal{Q}). \tag{5.6}$$

We obtain standard errors and 90% reverse-percentile confidence intervals by applying the same integration operator to each judge-cluster bootstrap MTE curve. These intervals summarize the average MTEs and their scalar difference. As a complementary whole-curve comparison, we test

$$H_0(\mathcal{Q}) : \mathrm{MTE}_g(p) = \mathrm{MTE}_{g'}(p), \quad \forall p \in \mathcal{Q}, \tag{5.7}$$

using a studentized bootstrap sup-t statistic.[29]

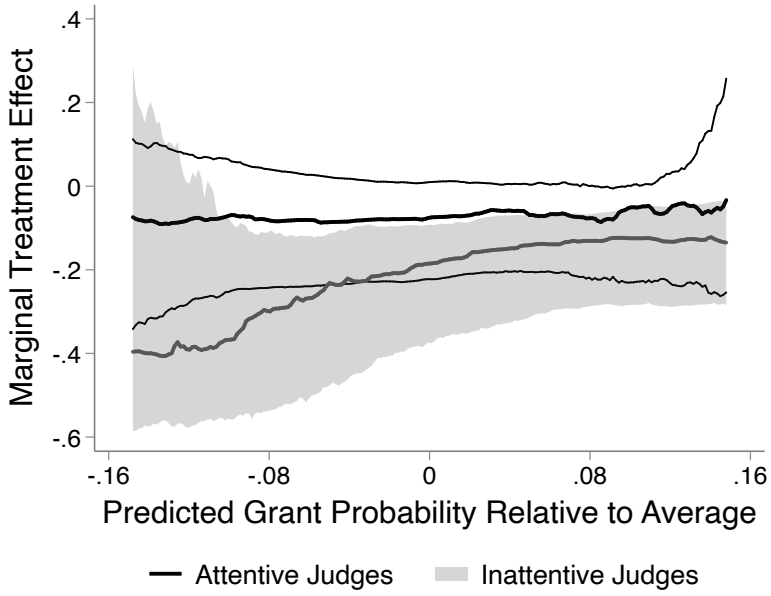
Figure 9 plots the estimated MTEs for attentive and inattentive judges over the shared grid of predicted grant probabilities. The estimated MTE for attentive judges remains close to zero across the support. For inattentive judges, the estimated MTE is more negative over much of the grid, with the largest separation between the two curves occurring at below-average predicted grant probabilities in the strict region. As predicted grant probability increases, the inattentive-judge MTE becomes less negative and the gap between the two curves narrows. These patterns mirror the reduced-form patterns in Figure 8b: the denial-conditioned reversal rate is nearly flat across leniency for attentive judges and declines steeply for inattentive judges, consistent with Lemma 5.1.

Table 2 quantifies the contrast in Figure 9. Over the full region, the average MTE is -0.072 for attentive judges and statistically insignificant. For inattentive judges, the average MTE is -0.219 and significantly different from zero. The

[28] This estimand summarizes the estimated MTE function rather than the distribution of p within each judge group. We refer to it as the *common-support average MTE*. Because lower predicted grant probability corresponds to lower judge leniency and a stricter decision threshold, we also report an average over the lower half of the grid, $\mathcal{P}^S = \{p \in \mathcal{P} : p \leq 0\}$, which we refer to as the strict region.

[29] The scalar comparison in Equation (5.6) summarizes the average difference between the two MTE functions. The sup-t test instead controls simultaneous inference over the entire function and is therefore more demanding. Failure to reject the global null indicates that the difference cannot be distinguished from zero over the entire specified region.

Figure 9: Marginal Treatment Effects by Grant-Rate Attentiveness



Notes: The figure plots the marginal treatment effect defined in Equation (5.4), separately for proceedings ruled by attentive and inattentive judges. The estimation sample consists of lower-court denials, and the outcome equals one when a denial is followed by an appeal grant and zero otherwise. Attentive and inattentive judges correspond to the top and bottom terciles of grant-rate response magnitude within ruling year, as described in Section 5.2.1. The horizontal axis is the mean-centered predicted lower-court grant probability; lower values correspond to lower judge leniency and stricter decision thresholds. The solid black line represents attentive judges, and the solid gray line represents inattentive judges. The dashed black lines and gray shaded area report pointwise 90% bootstrap percentile confidence intervals based on 500 replications resampled at the judge level.

attentive-minus-inattentive difference is 0.147, with a 90% confidence interval of $[-0.018, 0.349]$. The difference is 0.218 in the strict region, with a 90% confidence interval of $[0.009, 0.518]$. The whole-curve tests do not reject uniform equality of the two MTE functions over either the full region ($p = 0.425$) or the strict region ($p = 0.242$). Taken together, the estimates identify the strict region as the part of the predicted grant-probability distribution where average reversal sensitivity differs most clearly between attentive and inattentive judges. The threshold model interprets the flatter MTE for attentive judges as a decision threshold closer to c^*, where $R^D(c)$ is less sensitive to threshold changes. The estimates therefore link stronger responsiveness to closer distributional alignment with the BIA standard.

5.3.3 Appeal Filing and BIA Success: Marginal Treatment Effects

The denial-conditioned reversal outcome reflects two stages of the selected review process: whether a denied case is appealed and whether an appealed denial receives an appeal grant. We next ask whether the difference between attentive and inattentive judges appears in the appeal-filing margin, the BIA-success margin, or both. Specifically, the denial-conditioned reversal outcome can be written as

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

where $A_{ijt} = 1$ if denied case i is appealed and $A_{ijt} = 0$ otherwise, while $S_{ijt} = 1$ if the BIA grants the appeal, conditional on $A_{ijt} = 1$.[30] At the level of conditional means, differentiating with respect to p yields

$$\begin{aligned} \text{MTE}_Y(p) &= \frac{\partial}{\partial p} E\left[Y \mid p\right] \\ &= \underbrace{\underbrace{\text{MTE}_A(p)}_{\equiv \frac{\partial}{\partial p} E[A|p]} \cdot E\left[S \mid A{=}1, p\right]}_{\text{selection margin}} + \underbrace{E\left[A \mid p\right] \cdot \underbrace{\text{MTE}_S(p)}_{\equiv \frac{\partial}{\partial p} E[S|A=1,p]}}_{\text{success margin}}, \end{aligned} \tag{5.8}$$

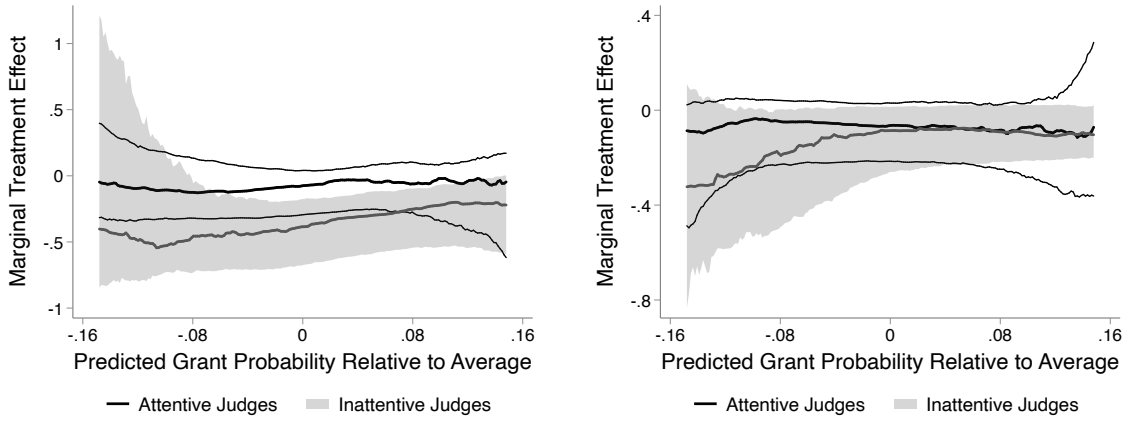
where all expectations are taken over lower-court denials. Equation (5.8) shows how marginal changes in appeal filing and BIA success enter the derivative of the denial-conditioned reversal probability. In the empirical analysis, we apply the residualized MTE procedure from Section 5.3.2 separately to each outcome on its relevant sample, using the same mean-centered common-support grid of predicted lower-court grant probabilities.

We begin with the selection margin. Figure 10a plots the estimated MTEs for the probability that a lower-court denial is appealed, separately for attentive and inattentive judges. For inattentive judges, the estimated MTE is negative across the displayed support and most negative in the strict region, indicating that the probability of appeal filing declines as judge leniency increases. In contrast, the attentive-judge MTE remains closer to zero across the support, indicating that appeal filing among denials changes much less with leniency. This flatter profile is consistent with a denial pool that changes less sharply with judge leniency when

[30] Following the disposition terminology in Section 2.3, this sample includes appealed lower-court denials classified as appeal grants or appeal denials. Neutral or other dispositions and non-appealed denials do not enter the BIA-success margin.

the decision threshold lies closer to the BIA standard.

Figure 10: Marginal Treatment Effects on Appeal Filing and BIA Success



(a) Selection Margin

(b) Success Margin

Notes: The figure plots margin-specific marginal treatment effects (MTEs) separately for attentive and inattentive judges, classified by grant-rate attentiveness as described in Section 5.2.1. Panel 10a uses lower-court denials and an outcome equal to one if the denial is appealed and zero otherwise. Panel 10b uses appealed lower-court denials with a BIA disposition classified as an appeal grant or appeal denial; neutral or other dispositions are excluded. Both panels apply the residualized MTE procedure described in Section 5.3.2 and use the same mean-centered common-support grid of predicted lower-court grant probabilities. The solid black line represents attentive judges, and the solid gray line represents inattentive judges. The dashed black lines and gray shaded areas report pointwise 90% bootstrap percentile confidence intervals based on 500 replications resampled at the judge level.

We next turn to the success margin, which measures the probability of an appeal grant among appealed lower-court denials. Figure 10b shows that the estimated MTE for inattentive judges is negative over much of the support and more negative at lower values of p, indicating that the appeal grant probability declines as judge leniency increases. The attentive-judge MTE remains closer to zero, although the separation between the two curves is smaller than on the selection margin. The success-margin MTE traces how the composition of appealed denials changes along the leniency distribution. This flatter profile for attentive judges is therefore consistent with an appealed-denial pool that changes less with leniency when the decision threshold lies closer to the BIA standard.

Columns (2) and (3) of Table 2 quantify these contrasts. On the selection margin, the attentive-minus-inattentive average-MTE difference is 0.288 over the full common support and 0.353 in the strict region; both are statistically significant at the 10% level. The corresponding success-margin differences are smaller and not statistically distinguishable from zero. The whole-curve tests do not reject equality

between the attentive- and inattentive-judge MTE functions for either margin over the full support or the strict region.

Taken together, these patterns are consistent with the model. For inattentive judges, greater threshold misalignment makes the composition of the denial pool more sensitive to marginal changes in leniency. For attentive judges, updating toward the BIA standard reduces this sensitivity, producing flatter MTEs. The clearest empirical contrast appears at selection into appeal, the first gate through which lower-court denials enter BIA review.

Table 2: Average Marginal Treatment Effects by Grant-Rate Attentiveness

	(1) Reversal Outcome	(2) Selection Margin	(3) Success Margin
Panel A: Common trimmed support, $p = [-0.1480, 0.1480]$			
$\overline{\text{MTE}}_{\text{att}}$	-0.0722	-0.0734	-0.0694
	[-0.1434, 0.0629]	[-0.2254, 0.1296]	[-0.1540, 0.0888]
$\overline{\text{MTE}}_{\text{inatt}}$	-0.2194	-0.3619	-0.1384
	[-0.3564, -0.0604]	[-0.7073, -0.1340]	[-0.2736, 0.0413]
$\overline{\text{MTE}}_{\text{att}} - \overline{\text{MTE}}_{\text{inatt}}$	0.1472	0.2884	0.0691
	[-0.0182, 0.3493]	[0.0034, 0.6896]	[-0.1168, 0.2656]
sup-t test	1.6540	1.7823	0.9596
global bootstrap p-value	0.4252	0.3254	0.8563
Panel B: Strict region, $p = [-0.1480, -0.0007]$			
$\overline{\text{MTE}}_{\text{att}}$	-0.0808	-0.1007	-0.0589
	[-0.1966, 0.0832]	[-0.3508, 0.1128]	[-0.1436, 0.1321]
$\overline{\text{MTE}}_{\text{inatt}}$	-0.2987	-0.4541	-0.1872
	[-0.5417, -0.1149]	[-1.0905, -0.2128]	[-0.3794, 0.0936]
$\overline{\text{MTE}}_{\text{att}} - \overline{\text{MTE}}_{\text{inatt}}$	0.2179	0.3533	0.1283
	[0.0094, 0.5184]	[0.0020, 0.9481]	[-0.1359, 0.3889]
sup-t test	1.6540	1.7823	0.9596
global bootstrap p-value	0.2415	0.1637	0.5868

Note: This table reports uniformly weighted average marginal treatment effects (MTEs) for proceedings ruled by attentive and inattentive judges, classified using grant-rate responsiveness as described in Section 5.2.1. For each judge group, the average MTE integrates the estimated MTE function over the indicated mean-centered predicted lower-court grant-probability region and divides by the width of that region. Panel A uses the full common trimmed support, while Panel B uses the lower half of that support, corresponding to the strict region emphasized by the model in Section 5.1. Column (1) uses the denial-conditioned reversal outcome, as in Figure 9. Column (2) uses the selection-margin outcome, as in Figure 10a. Column (3) uses the success-margin outcome, as in Figure 10b. Brackets report 90% reverse-percentile confidence intervals based on 500 bootstrap replications clustered at the judge level. The attentive-minus-inattentive row reports the scalar difference between the two average MTEs. The sup-t statistic and global bootstrap p-value test equality of the two MTE functions at every point over the panel-specific support.

Appendix Figures A.31 and A.32 compare MTE curves using changes in early

predictability and the number of hearings per proceeding after surprise reversals as proxies for attentiveness in the hearing process. The hearing-count classification points in the same direction as the grant-rate result, with attentive judges exhibiting MTEs closer to zero on average, whereas the early-predictability classification does not yield the same differentiation. Both differences are imprecisely estimated (Appendix Tables A.9 and A.10). Appendix Figures A.33 and A.34 report the corresponding selection- and success-margin MTEs, which also provide limited evidence of systematic group separation. Taken together, grant-rate responsiveness provides the clearest empirical link between attentiveness and distributional alignment with the BIA standard, while early-predictability and hearing-count responsiveness capture complementary dimensions of attentiveness in the lower-court hearing process.

By construction, grant-rate attentiveness is proxied by absolute responsiveness, although immigration judges increase grant rates on average after surprise reversals. We therefore separate attentive judges by the direction of their grant-rate response and estimate MTEs for the two groups. Appendix Figure A.35 shows that the MTE curves for positive- and negative-reacting attentive judges are not systematically distinguishable.[31] This result supports response magnitude, rather than the direction of the grant-rate change, as the dimension most closely associated with alignment.

5.3.4 Case Heterogeneity

Whether appellate feedback translates into aligned future decisions may also depend on the types of cases in which it is generated and later applied. We examine two dimensions of case heterogeneity: procedural type and applicants' countries of origin.

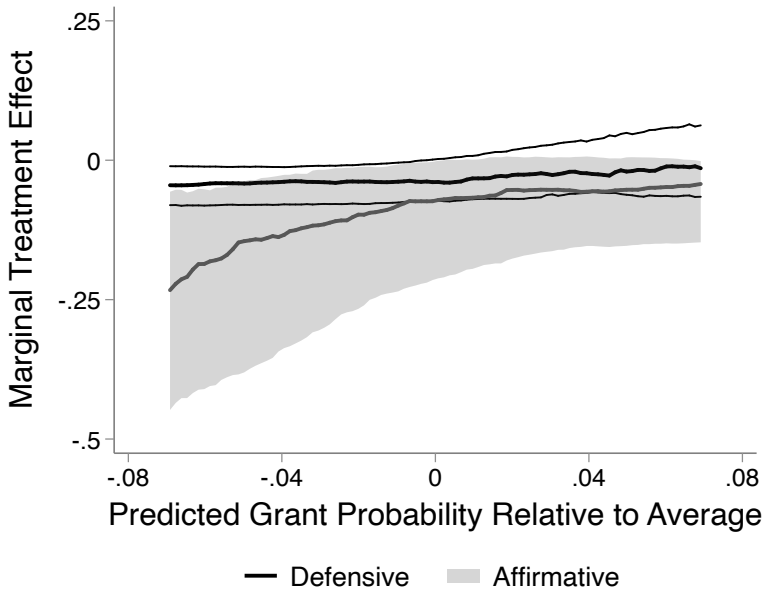
Procedural heterogeneity. We first test whether distributional alignment differs between *defensive* and *affirmative* proceedings, the two major procedural categories in our data. Defensive claims arise in removal proceedings and often involve higher-stakes consequences for applicants, more developed evidentiary records,

[31] The average MTE difference between positive- and negative-reacting attentive judges on the common support is -0.086 with the 90% confidence interval $[-0.544, 0.487]$; the difference on the strict region is -0.087 with the 90% confidence interval $[-0.456, 0.660]$.

and more structured hearings. Affirmative claims are filed voluntarily by asylum seekers not yet in removal proceedings. These differences may affect both which denials are appealed and how readily BIA feedback applies to future decisions: feedback from defensive proceedings may be more salient and transferable across similar cases than feedback from affirmative proceedings due to the higher stakes, more developed records, and more structured hearings. We therefore hypothesize that lower-court decisions in defensive proceedings are more closely aligned with the BIA standard.

The MTEs in Figure 11 support this hypothesis. The estimated MTE is close to zero for defensive proceedings and more negative for affirmative proceedings. The defensive-minus-affirmative difference is 0.067 (S.E. 0.068) over the common trimmed support and 0.104 (S.E. 0.093) in the strict region. The estimates have the predicted sign in both regions, although neither difference is statistically distinguishable from zero.

Figure 11: Marginal Treatment Effect, Procedural Heterogeneity

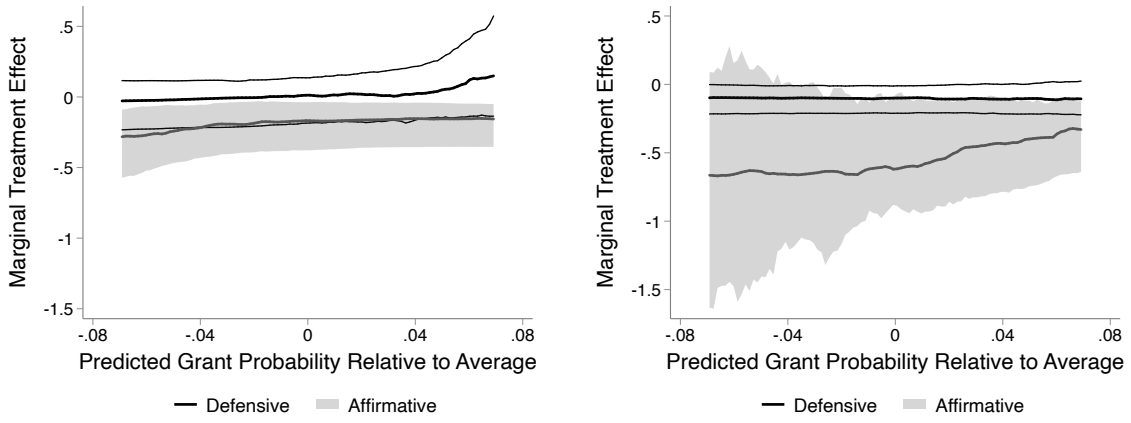


Notes: The figure plots the marginal treatment effects (MTEs) of lower-court decisions on lower-court reversals separately for affirmative versus defensive cases. The MTEs are estimated following the procedure outlined in Section 5.3. The black solid line shows the MTEs estimated on the subsample of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

The pooled comparison does not reveal whether procedural differences are con-

centrated among attentive or inattentive judges. Figure 12 therefore plots the MTEs for defensive and affirmative proceedings separately for the two response groups. Among attentive judges, the MTE curves are comparatively flat in both categories. The defensive-minus-affirmative average-MTE difference is 0.213 over the common support, with a 90% confidence interval of $[-0.041, 0.456]$, and 0.212 in the strict region, with a 90% confidence interval of $[-0.040, 0.492]$. The procedural contrast is substantially larger among inattentive judges. Their defensive-proceeding MTE remains relatively flat, whereas the affirmative-proceeding MTE is sharply negative, particularly in the strict region. Table A.11 reports a defensive-minus-affirmative average-MTE difference of 0.443 over the common support, with a 90% confidence interval of $[0.147, 0.822]$. The difference increases to 0.547 in the strict region with a 90% confidence interval of $[-0.026, 1.062]$. These differences therefore suggest that procedural heterogeneity is concentrated among inattentive judges. This pattern is consistent with defensive proceedings making BIA feedback easier to translate into future decisions, even among judges whose overall response to surprise reversals is low.

Figure 12: Marginal Treatment Effects by Procedural Type and Attentiveness



(a) Attentive Judges

(b) Inattentive Judges

Notes: The figure plots the marginal treatment effect defined in Equation (5.4) separately for defensive and affirmative proceedings. Panel 12a uses proceedings ruled by attentive judges, and Panel 12b uses proceedings ruled by inattentive judges. The outcome equals one when a denial is followed by an appeal grant and zero otherwise. Attentive and inattentive judges correspond to the top and bottom terciles of grant-rate response magnitude within ruling year, as described in Section 5.2.1. Both panels use the common mean-centered grid of predicted lower-court grant probabilities for the procedural comparison. In each panel, the solid black line represents defensive proceedings, and the solid gray line represents affirmative proceedings. The dashed black lines and gray shaded area report pointwise 90% bootstrap percentile confidence intervals based on 500 replications resampled at the judge level.

Source-country heterogeneity. We next examine whether distributional alignment varies across applicants' countries of origin. We consider three classifications: Latin American versus other origins, recent armed conflict versus no recent armed conflict, and lower-income versus other origins. The main comparisons pool judges across response groups and provide supplementary evidence on substantive case heterogeneity.

First, we compare proceedings involving applicants from Latin American countries with proceedings involving applicants from other countries.[32] Latin American origins account for 48.10% of proceedings in the sample. Their average lower-court grant rate is 17.67%, compared with 52.43% for other origins. Appendix Figure A.36 presents the corresponding reduced-form and MTE comparisons. Over the common support, the average MTE is -0.006 for Latin American origins and -0.156 for other origins. The Latin-America-minus-other difference is 0.150, with a 90% confidence interval of $[-0.028, 0.332]$. The difference is similar in the strict region at 0.170, with a 90% confidence interval of $[-0.003, 0.373]$. The point estimates therefore indicate closer distributional alignment for Latin American origins. The subgroup estimates point in the same direction: Latin American MTEs are closer to zero for both attentive and inattentive judges, although the between-origin differences are imprecisely estimated within each response group. Appendix Tables A.12 and A.13 report the complete estimates.

Second, we classify an origin as recently conflict-affected if the UCDP/PRIO Armed Conflict Dataset records at least one state-based armed conflict during the lower-court decision year or either of the two preceding years (Gleditsch et al., 2002; Davies, Pettersson and Öberg, 2026).[33] This group accounts for 32.72% of proceedings and has an average lower-court grant rate of 40.4%. Appendix Figure A.37 shows a more negative pooled MTE for conflict-affected origins, indicating less distributional alignment. The conflict-minus-other average-MTE difference is -0.075 over the common support, with a 90% confidence interval of $[-0.143, 0.007]$, and -0.084 in the strict region, with an interval of $[-0.168, 0.005]$. This directional

[32] We use the United Nations subregion Latin American and Caribbean Group (GRULAC), which consists of 33 member states. See https://www.un.org/dgacm/en/content/regional-groups.

[33] We use version 26.1, accessed on July 1, 2026, from https://ucdp.uu.se/downloads/ (Pettersson, 2026). UCDP defines a state-based armed conflict as a contested incompatibility over government or territory involving armed force between two parties, at least one of which is a state government, that results in at least 25 battle-related deaths during a calendar year.

difference is concentrated among inattentive judges but remains imprecisely estimated.

Finally, we construct a time-varying lower-income-origin indicator based on GDP per capita at purchasing-power parity in constant 2021 international dollars. The indicator equals one when the applicant's country of origin lies in the bottom half of the available country distribution for that year.[34] Lower-income origins account for 83.47% of proceedings and have an average lower-court grant rate of 35.26%. Appendix Figure A.38 shows little separation between the two MTE curves. The lower-income-minus-other average-MTE difference is 0.032 over the common support, with a 90% confidence interval of $[-0.032, 0.107]$, and 0.022 in the strict region, with an interval of $[-0.046, 0.101]$. This pattern is consistent across attentive and inattentive judges. Unlike the Latin American and recent-conflict comparisons, this result shows that source-country heterogeneity does not follow a simple country-income gradient.

Taken together, these results show that the relationship between attentiveness and distributional alignment varies systematically across case types. For defensive proceedings, even inattentive judges exhibit relatively flat MTEs, consistent with the higher stakes and more developed procedural records making appellate feedback more salient and easier to apply. The difference between attentive and inattentive judges is substantially larger for affirmative proceedings. The closer alignment for Latin American origins across both judge groups may reflect greater external scrutiny, which can make these cases more salient to immigration judges. The weaker alignment for recent-conflict origins, particularly among inattentive judges, may instead reflect abrupt conflict-related changes in caseload volume or composition that make prior BIA feedback less readily transferable. These interpretations are consistent with the observed patterns, although the analysis does not separately identify either mechanism. The absence of meaningful separation by country income provides an informative benchmark: not every country classification produces different MTE curves, making the Latin American and recent-conflict results less likely to reflect arbitrary country groupings. More generally,

[34] We use the World Bank's World Development Indicators from 1990 onward whenever available (World Bank, 2026), Penn World Table version 11.0 for 1985–1989 (Feenstra, Inklaar and Timmer, 2025), and Penn World Table (Feenstra, Inklaar and Timmer, 2015) as a fallback after 1990. We exclude regional and income-group aggregates; observations without an available GDP estimate remain missing.

attentive judges exhibit MTEs closer to zero across these classifications, reinforcing the main finding that greater responsiveness to appellate feedback is associated with closer distributional alignment with the BIA standard.

In summary, the MTE results provide a structural counterpart to the reduced-form evidence in Section 5.3.1. Judges who respond more strongly to surprise reversals exhibit flatter MTEs and closer distributional alignment with the BIA standard, particularly in the strict region. This relationship appears through both the appeal-filing and BIA-success margins. The case-heterogeneity results further show that alignment depends on both judges' responsiveness and the settings in which appellate feedback is generated and applied. Together, these findings show how attention shapes whether hierarchical review informs future lower-court decisions.

6 Conclusion

Hierarchical review serves two organizational functions: it resolves or reconsiders the case before the reviewing authority, and it can inform the decentralized expert's future decisions. In this paper, we study the feedback function of review in U.S. asylum adjudication, where formal feedback is selectively generated through an immigration judge's denial, the applicant's decision to appeal, and a BIA disposition. We distinguish appellate signals by diagnosticity, estimate immigration judges' behavioral responses, and examine whether judges with greater response magnitude exhibit closer distributional alignment with the BIA standard.

Using linked EOIR and BIA administrative records, we construct two rolling prediction models with distinct empirical roles. The lower-court model measures the predictability of final rulings from information fixed before the hearing process. The BIA model provides an observable ex ante reversal-risk benchmark that classifies realized reversals into expected and surprise groups by diagnostic content. Both groups share the same formal appellate direction, but surprise reversals are less predictable from pre-resolution observables and therefore carry greater diagnostic content relative to the model-based benchmark. The event-study estimates show a stronger response to these more diagnostic signals: relative to the seven days preceding the recorded BIA decision, immigration judges' grant rates from the decision date through six days afterward rise by 1.57 percentage points

more after surprise reversals than after expected reversals. The stronger grant-rate response is not accompanied by a broad decline in our observable proxies for lower-court hearing diligence. Early predictability shows no systematic post-event increase, hearing records per proceeding rise modestly in the first post-event week, and the other procedural measures show no systematic immediate change. Grant-rate responses vary substantially across immigration judges in both magnitude and direction, and we use the absolute value of each judge-specific estimate as our primary measure of behavioral responsiveness.

To connect behavioral responsiveness to ruling standard alignment, we develop a threshold model in which a judge-specific effective updating rate governs how appellate feedback moves future decision thresholds toward the BIA standard. We classify judges as high-response or low-response based on their grant-rate response magnitudes, and interpret greater responsiveness as more effective attention and updating. We then estimate marginal treatment effects (MTEs) for the probability that a lower-court denial is followed by a BIA grant. The estimated MTE for high-response judges remains close to zero, while the MTE for low-response judges is more negative over much of the predicted grant-probability distribution, with the clearest average difference at below-average probabilities where decision thresholds are stricter. The threshold model interprets this flatter MTE among high-response judges as closer distributional alignment with the BIA standard. The MTE curves for positive- and negative-reacting judges within the high-response group are not systematically distinguishable, supporting response magnitude, rather than direction, as the dimension most closely associated with alignment. We next estimate MTEs for both review margins: appeal filing among lower-court denials and BIA success conditional on appeal. The same response-group ordering appears on both margins but is clearest at appeal filing; the BIA-success difference is smaller and not statistically distinguishable. These estimates characterize two distinct stages of selectively generated review.

The relationship between responsiveness and alignment also varies across procedural settings. The defensive-versus-affirmative contrast is substantially larger among low-response judges, whose MTE is relatively flat in defensive proceedings but sharply negative in affirmative proceedings, particularly where decision thresholds are stricter. Among high-response judges, the MTE is comparatively flat in both. This pattern is consistent with the more developed records and structured

hearings of defensive proceedings making BIA feedback easier to apply across similar cases. Point estimates from supplementary source-country comparisons indicate closer distributional alignment for Latin American origins and weaker alignment for recently conflict-affected origins, but little separation by country income. The absence of an income gradient shows that not every country classification generates a distinct MTE pattern.

Taken together, the findings show that the teaching function of hierarchical review depends on both the production and use of feedback. Review design shapes which decisions can generate formal feedback, signal diagnosticity shapes how strongly recipients respond, and response magnitude is associated with closer distributional alignment with the higher-level standard. Hierarchical review therefore serves not only to evaluate the current decision but also as an information technology for future decentralized decisions.

Our analysis evaluates distributional alignment with the BIA standard on the denial-appeal margin. Closer alignment means that decentralized decisions track the organization's reviewing standard more consistently, an important dimension of organizational performance; whether the BIA standard is socially optimal is a separate welfare question. Under the threshold model, the relationship between response magnitude and alignment shows how attention to diagnostic feedback can strengthen consistency within the review hierarchy. Future work can connect organizational alignment to broader legal and welfare outcomes and examine how formal review interacts with experiential and collective learning. More broadly, by separating signal diagnosticity from recipient responsiveness, our framework provides a general approach for studying how selectively generated review informs future decisions across a wide range of expert settings.

References

Aghion, Philippe, and Jean Tirole. 1997. "Formal and real authority in organizations." *Journal of political economy*, 105(1): 1–29.

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.

Argote, Linda, and Ella Miron-Spektor. 2011. "Organizational Learning: From Experience to Knowledge." *Organization Science*, 22(5): 1123–1137.

Argote, Linda, Sunkee Lee, and Jisoo Park. 2021. "Organizational learning processes and outcomes: Major findings and future research directions." *Management science*, 67(9): 5399–5429.

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 H. C. Sant'Anna. 2026. "Difference-in-Differences Designs: A Practitioner's Guide." *Journal of Economic Literature*, 64(2): 498–557.

Beim, Deborah. 2017. "Learning in the judicial hierarchy." *The Journal of Politics*, 79(2): 591–604.

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.

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

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.

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.

Davies, Shawn, Therése Pettersson, and Magnus Öberg. 2026. "Organized Violence 1989–2025, and Violent Political Protests." *Journal of Peace Research*, 63(4): 705–723.

Denrell, Jerker, and James G. March. 2001. "Adaptation as Information Restriction: The Hot Stove Effect." *Organization Science*, 12(5): 523–538.

Dessein, Wouter. 2002. "Authority and Communication in Organizations." *Review of Economic Studies*, 69(4): 811–838.

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.

Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer. 2015. "The Next Generation of the Penn World Table." *American Economic Review*, 105(10): 3150–3182.

Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer. 2025. "Penn World Table, Version 11.0." Released October 7, 2025; accessed August 1, 2026.

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.

Garicano, Luis. 2000. "Hierarchies and the Organization of Knowledge in Production." *Journal of political economy*, 108(5): 874–904.

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

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand. 2002. "Armed Conflict 1946–2001: A New Dataset." *Journal of Peace Research*, 39(5): 615–637.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2026. "Leniency Designs: An Operator's Manual." *Journal of Economic Perspectives*, 40(3): 213–240.

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." *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.

Joseph, John, Ronald Klingebiel, and Alex James Wilson. 2016. "Organizational structure and performance feedback: Centralization, aspirations, and termination decisions." *Organization Science*, 27(5): 1065–1083.

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): 579–606.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." 275–284. New York, NY, USA:Association for Computing Machinery.

Levitt, Barbara, and James G. March. 1988. "Organizational Learning." *Annual Review of Sociology*, 14: 319–340.

Madsen, Peter M., and Vinit Desai. 2010. "Failing to Learn? The Effects of Failure and Success on Organizational Learning in the Global Orbital Launch Vehicle Industry." *Academy of Management Journal*, 53(3): 451–476.

Ocasio, William. 1997. "Towards an Attention-Based View of the Firm." *Strategic Management Journal*, 18(S1): 187–206.

Pettersson, Therése. 2026. "UCDP/PRIO Armed Conflict Dataset Codebook, Version 26.1." Version 26.1.

Prendergast, Canice. 2003. "The limits of bureaucratic efficiency." *Journal of Political Economy*, 111(5): 929–958.

Rahmandad, Hazhir. 2008. "Effect of delays on complexity of organizational learning." *Management Science*, 54(7): 1297–1312.

Ramji-Nogales, Jaya, Andrew I Schoenholtz, and Philip G Schrag. 2007. "Refugee roulette: Disparities in asylum adjudication." *Stan. L. Rev.*, 60: 295–411.

Sah, Raaj Kumar, and Joseph E. Stiglitz. 1986. "The Architecture of Economic Systems: Hierarchies and Polyarchies." *American Economic Review*, 76(4): 716–727.

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

Shavell, Steven. 1995. "The Appeals Process as a Means of Error Correction." *Journal of Legal Studies*, 24(2): 379–426.

Shavell, Steven. 2006. "The appeals process and adjudicator incentives." *The Journal of Legal Studies*, 35(1): 1–29.

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.

World Bank. 2026. "World Development Indicators: GDP per capita, PPP (constant 2021 international $)." Indicator NY.GDP.PCAP.PP.KD; accessed August 1, 2026.

A Figures and Tables

Table A.1: Top Countries of Origin in EOIR 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

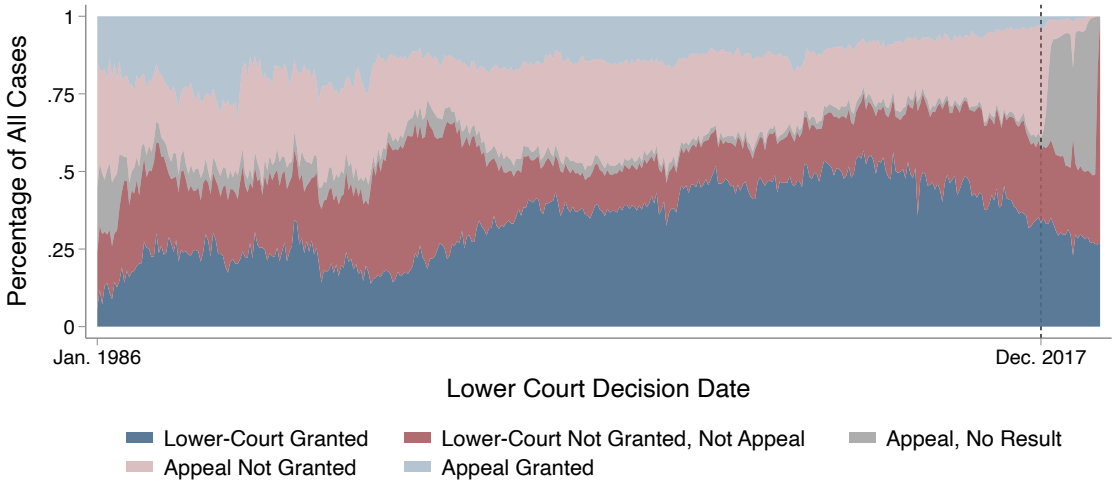
Notes: Counts report unique proceedings in the cleaned EOIR proceeding sample described in Section 2.3, covering lower-court decisions from 1986 through 2019.

Table A.2: Top U.S. States by Number of EOIR Proceedings, 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

Notes: Counts report unique proceedings in the cleaned EOIR proceeding sample described in Section 2.3, covering lower-court decisions from 1986 through 2019. Hearing state is defined using the EOIR hearing-location code.

Figure A.1: Monthly Distribution of Lower-Court and Appellate Outcomes, 1986-2019



Notes: The figure reports monthly outcome shares among proceedings in the cleaned EOIR proceeding sample described in Section 2.3. The denominator in each month is the number of proceedings decided by immigration judges in that month. The five mutually exclusive categories are: (i) a first-instance grant; (ii) a first-instance denial without a matched appeal; (iii) an appealed denial without a coded appeal-grant or appeal-denial outcome in the variable used to construct the figure; (iv) an appeal denial; and (v) an appeal grant. Appeal grant and appeal denial are shorthand for the positive and negative BIA disposition categories defined in Appendix Table A.3. The residual category labeled "Appeal, No Result" in the graphic includes other dispositions and unresolved appeals.

Table A.3: BIA Disposition Categories

Category	BIA 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 Affirmance/VD
	Summary Dismiss
	Summary Dismissal (O) Other
	Summary Dismissal (a) inad reason on appeal
	Summary Dismissal - Both (a) & (e)
Neutral or Other	Summary Dismissal (e) no brief
	Administrative Return
	Dismiss as Untimely
	Lacks Jurisdiction - BIA
	Moot Bond
	Withdrawal of Appeal
	Continued
	Other
	Other Soriano

Notes: This table lists 27 distinct BIA decision types and groups them into positive, negative, and neutral or other disposition categories using EOIR BIA decision codes. *Appeal grant* and *appeal denial* are shorthand for the positive and negative disposition categories, respectively. An appeal grant may include a remand or another positive BIA disposition and therefore does not necessarily imply a final grant of asylum. An appeal denial may include an affirmance, dismissal, or another negative BIA disposition and does not necessarily imply a denial on the merits. Neutral or other dispositions are excluded from the binary BIA prediction sample.

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

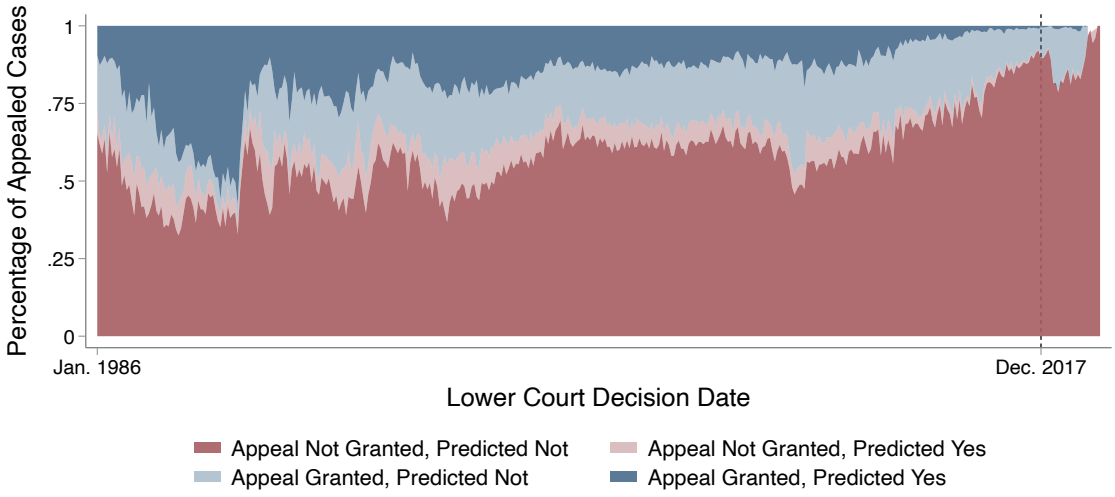
Notes: This table reports the mean of selected judge characteristics based on 698 EOIR judges.

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

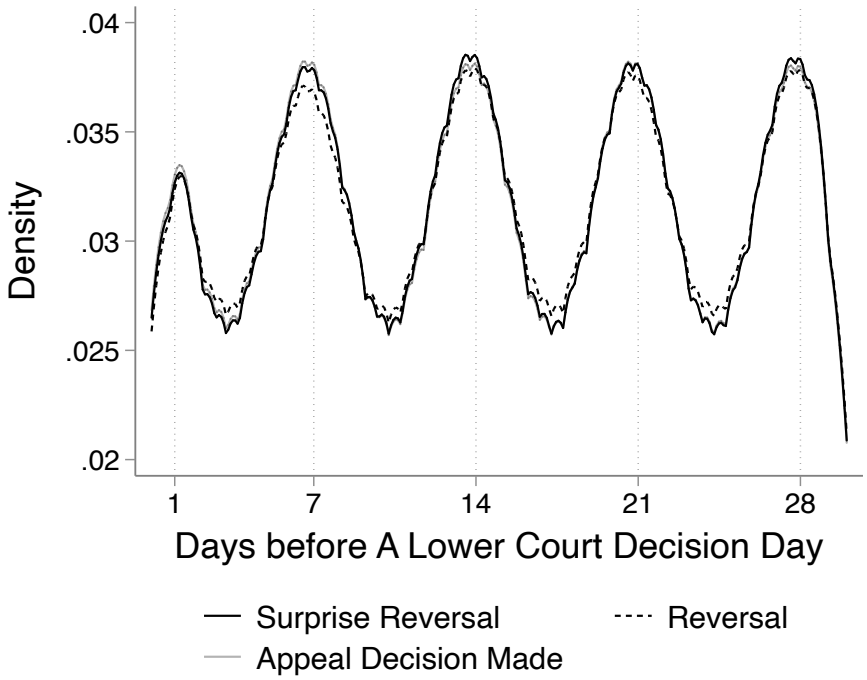
Notes: Each row reports the predictive performance of one of the five early-predictability models described in Section 3.1. The reported metrics are the area under the receiver operating characteristic curve (ROC AUC), classification accuracy, and log loss. All five specifications use random-forest estimators and are evaluated using the sequential prediction design described in Appendix Section B.1. Hyperparameters are selected separately through three-fold stratified cross-validation on the aggregate sample.

Figure A.2: Monthly Distribution of Realized and Predicted BIA Outcomes



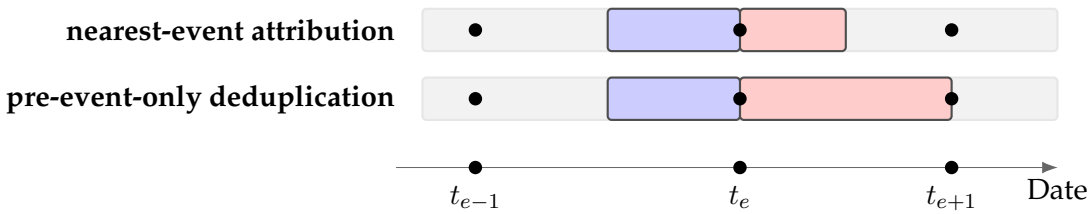
Notes: The figure reports monthly shares of realized and predicted outcomes among appeal records with a positive or negative BIA disposition and a valid classifier assignment. The denominator in each month is the number of these appeal records linked to lower-court proceedings decided in that month. The four mutually exclusive categories are: (i) a realized appeal denial predicted as an appeal denial, or a true negative; (ii) a realized appeal denial predicted as an appeal grant, or a false positive; (iii) a realized appeal grant predicted as an appeal denial, or a false negative; and (iv) a realized appeal grant predicted as an appeal grant, or a true positive. Appeal grant and appeal denial are shorthand for the positive and negative BIA disposition categories defined in Appendix Table A.3. The horizontal axis reports the month of the linked lower-court decision.

Figure A.3: Timing of Recorded BIA Decisions before Lower-Court Rulings



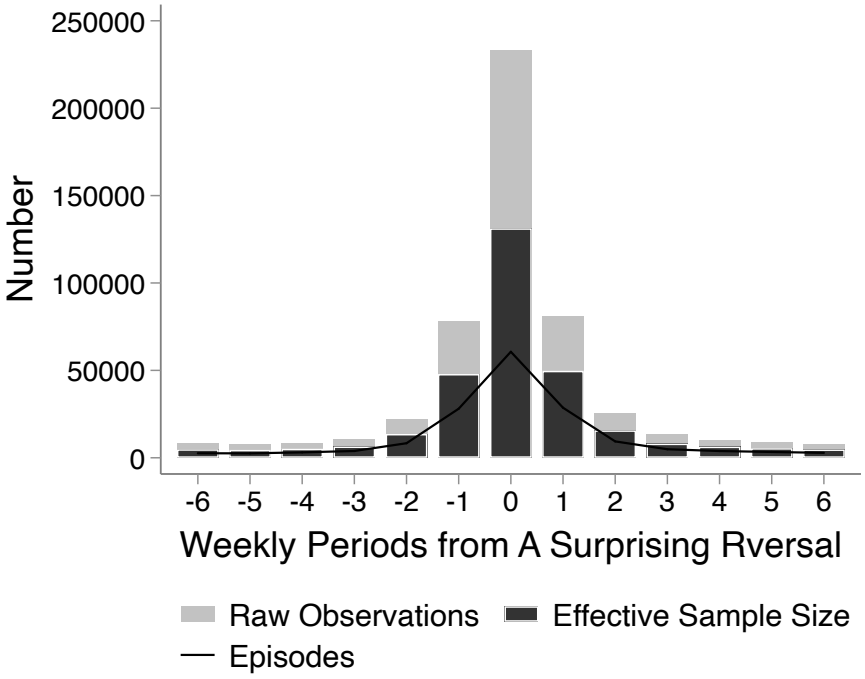
Notes: The figure plots kernel-density estimates of the number of days between a lower-court ruling and recorded BIA decisions involving the same immigration judge during the preceding 30 days. The three series include all linked appeal-grant and appeal-denial BIA decisions, realized reversals, and surprise reversals. A realized reversal is an appeal grant following the immigration judge's lower-court denial, and a surprise reversal is a realized reversal that the classifier predicted would be denied. Densities use an Epanechnikov kernel with bandwidth one. The recurring seven-day pattern motivates the seven-day event-time periods used in Section 4.1.

Figure A.4: Event-Study Sample-Trimming Rules



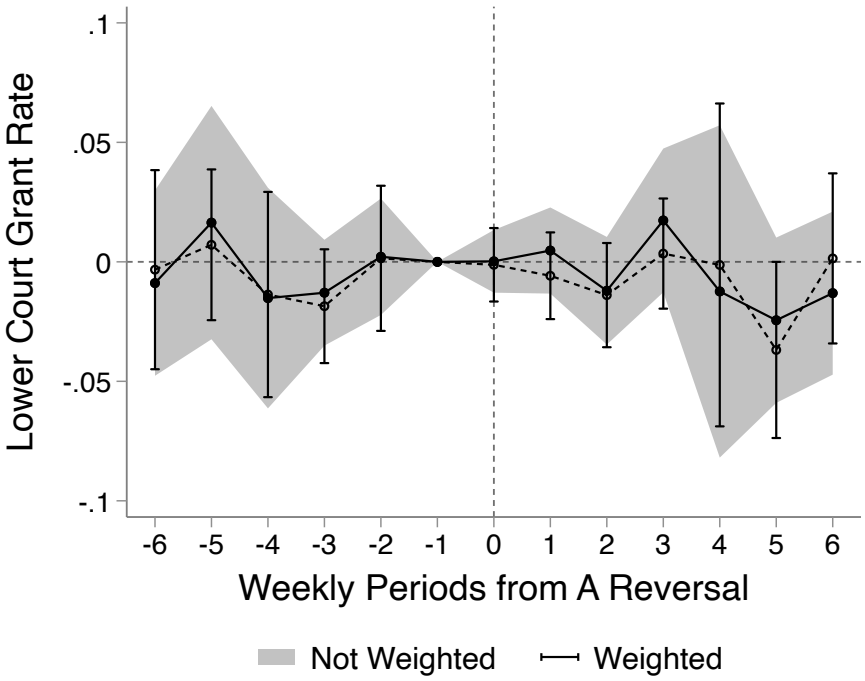
Notes: The figure provides a stylized illustration of the nearest-event-attribution and pre-event-only-deduplication rules described in Section 4.1. Dots mark the preceding event, focal event, and subsequent event. Blue and red regions denote retained pre-event and post-event linkages, respectively, while gray regions denote excluded linkages. The implemented rules operate on weekly event-time indices and retain tied linkages at the relevant minimum absolute index.

Figure A.5: Event-Time Support in the Feedback-Diagnosticity Sample



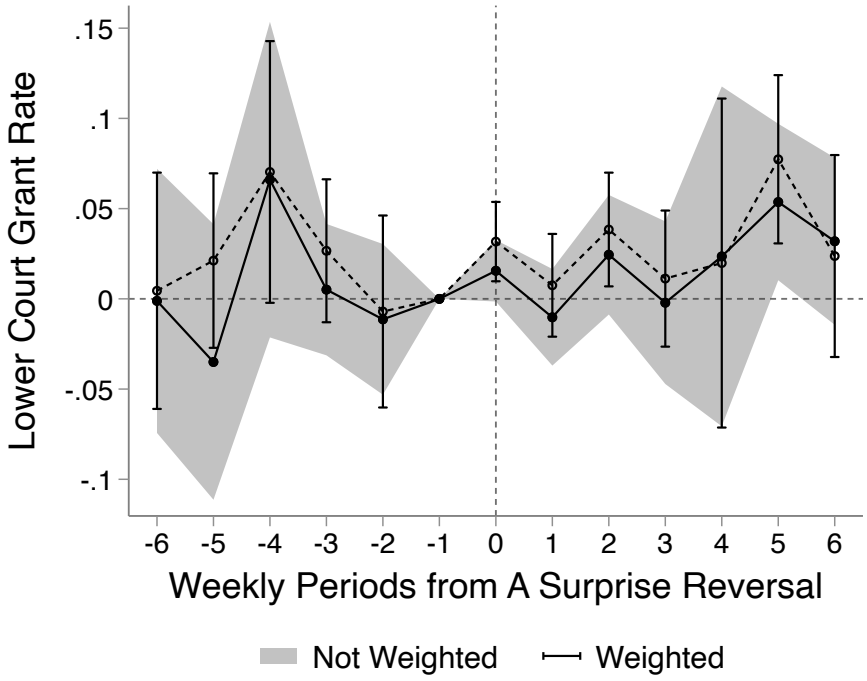
Notes: The figure summarizes support across the unbinned event periods in the feedback-diagnosticity sample after nearest-event attribution. The sample contains realized reversals with valid classifier assignments, classified as expected or surprise reversals. Raw observations are the retained case-event-period observations at each event time k. Effective sample size is calculated from the event-equal weights as $\left(\sum_{i,e} w_{ie}(k)\right)^2 / \sum_{i,e} w_{ie}(k)^2$. The series labeled "Episodes" in the graphic reports the number of distinct BIA events represented at each k. Section 4.1 defines the sample, attribution rule, and weights.

Figure A.6: Lower-Court Grant-Rate Responses: Appeal Grants versus Appeal Denials, Unbinned



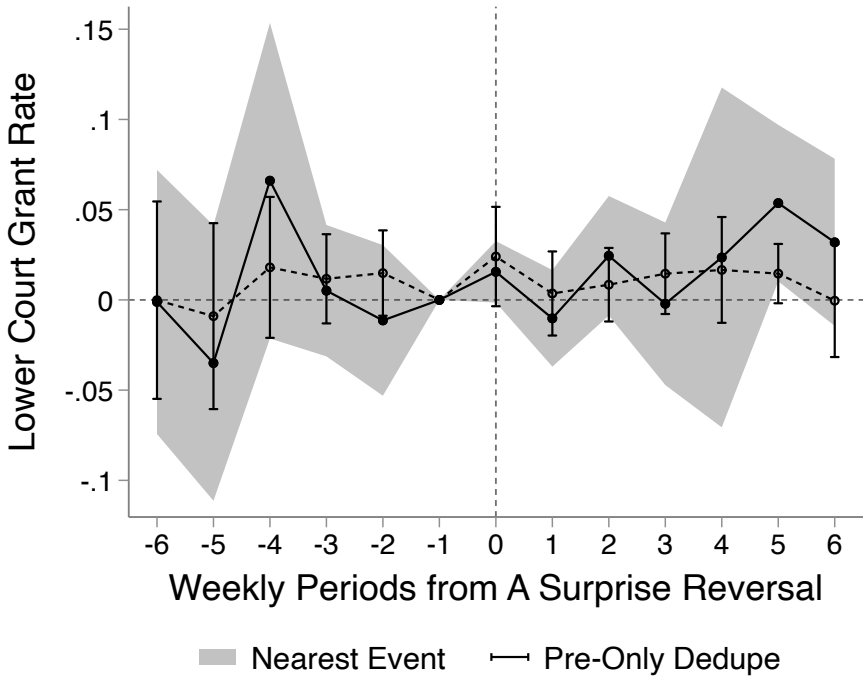
Notes: The figure plots $\hat{\beta}_k$ from Specification (4.1), estimated separately for each seven-day period $k \in \{-6, \ldots, 6\}$, with period $k = -1$ omitted. Each coefficient measures how the change in lower-court grant rates around appeal-grant events differs from the corresponding change around appeal-denial events, relative to the omitted period. The general BIA-outcome event sample contains filtered and deduplicated BIA appeal records with an appeal-grant or appeal-denial disposition, a valid recorded BIA decision date, and a link to the immigration judge who issued the corresponding lower-court denial. Lower-court proceedings are linked to events using nearest-event attribution. Period $k = 0$ begins on the recorded BIA decision date and ends six days later. All specifications include BIA event fixed effects, BIA decision-year-by-day-of-week fixed effects, and BIA decision-month-by-day-of-week fixed effects. Standard errors are clustered by immigration judge. Solid dots with shaded pointwise 95% confidence intervals report the unweighted estimates; open circles with capped pointwise 95% confidence intervals report the event-equal-weighted estimates. Section 4.1 defines the event sample, attribution rule, and weights.

Figure A.7: Lower-Court Grant-Rate Responses: Surprise versus Expected Reversals, Unbinned



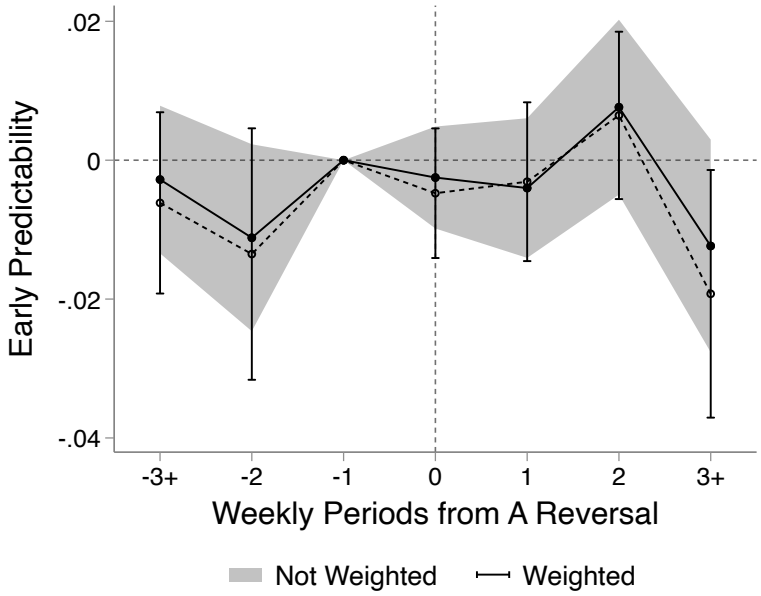
Notes: The figure plots $\hat{b}_k$ from Specification (4.2), estimated separately for each seven-day period $k \in \{-6, \ldots, 6\}$, with period $k = -1$ omitted. Each coefficient measures how the change in lower-court grant rates around surprise reversals differs from the corresponding change around expected reversals, relative to the omitted period. The feedback-diagnosticity event sample contains realized reversals with valid classifier assignments: an expected reversal is one for which the classifier predicts an appeal grant, and a surprise reversal is one for which the classifier predicts an appeal denial. Lower-court proceedings are linked to events using the nearest-event-attribution rule described in Section 4.1. Period $k = 0$ begins on the recorded BIA decision date and ends six days later. All specifications include BIA event fixed effects, BIA decision-year-by-day-of-week fixed effects, and BIA decision-month-by-day-of-week fixed effects. Standard errors are clustered by immigration judge. Solid dots with shaded pointwise 95% confidence intervals report the unweighted estimates; open circles with capped pointwise 95% confidence intervals report the event-equal-weighted estimates.

Figure A.8: Lower-Court Grant-Rate Responses under Alternative Trimming Rules: Surprise versus Expected Reversals



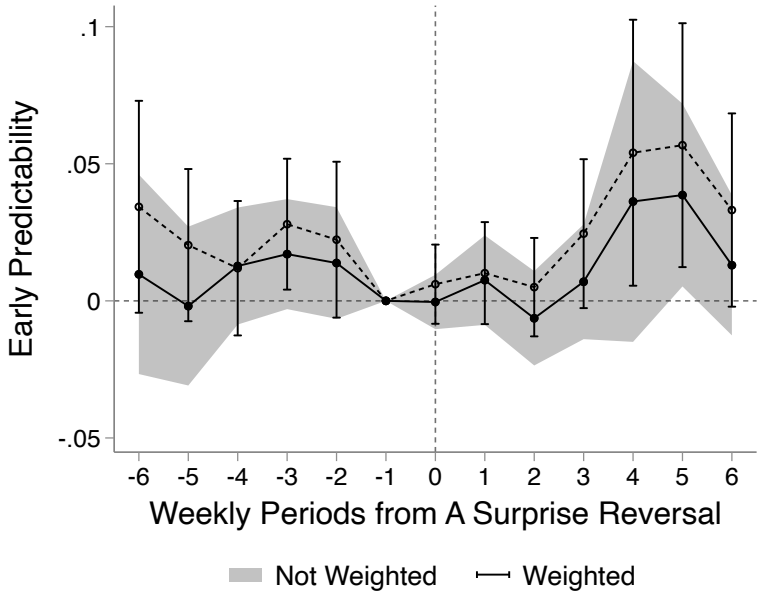
Notes: The figure plots unweighted estimates of $\hat{b}_k$ from Specification (4.2), estimated separately for each seven-day period $k \in \{-6, \ldots, 6\}$, with period $k = -1$ omitted. Each coefficient measures how the change in lower-court grant rates around surprise reversals differs from the corresponding change around expected reversals, relative to the omitted period. The feedback-diagnosticity event sample contains realized reversals with valid classifier assignments: an expected reversal is one for which the classifier predicts an appeal grant, and a surprise reversal is one for which the classifier predicts an appeal denial. The two series apply the nearest-event-attribution and pre-event-only-deduplication rules described in Section 4.1. Period $k = 0$ begins on the recorded BIA decision date and ends six days later. All specifications include BIA event fixed effects, BIA decision-year-by-day-of-week fixed effects, and BIA decision-month-by-day-of-week fixed effects. Standard errors are clustered by immigration judge. Solid dots with shaded pointwise 95% confidence intervals report estimates under nearest-event attribution; open circles with capped pointwise 95% confidence intervals report estimates under pre-event-only deduplication.

Figure A.9: Lower-Court Early-Predictability Responses: Appeal Grants versus Appeal Denials



Notes: The figure plots $\hat{\beta}_\kappa$ from Specification (4.1) using the continuous early-predictability score defined in Section 4.3 as the outcome. Each coefficient measures how the change in early predictability around appeal-grant events differs from the corresponding change around appeal-denial events, relative to the omitted period $k = -1$. The general BIA-outcome event sample contains filtered and deduplicated BIA appeal records with an appeal-grant or appeal-denial disposition, a valid recorded BIA decision date, and a link to the immigration judge who issued the corresponding lower-court denial. The estimation further requires a valid early-predictability score for the linked lower-court proceeding. Lower-court proceedings are linked to events using nearest-event attribution. The x-axis indexes binned seven-day periods around the recorded BIA decision date; period $k = 0$ begins on that date and ends six days later. All specifications include BIA event fixed effects, BIA decision-year-by-day-of-week fixed effects, and BIA decision-month-by-day-of-week fixed effects. Standard errors are clustered by immigration judge. Solid dots with shaded pointwise 95% confidence intervals report the unweighted estimates; open circles with capped pointwise 95% confidence intervals report the event-equal-weighted estimates. Section 4.1 describes the linkage and weighting procedures.

Figure A.10: Lower-Court Early-Predictability Responses: Surprise versus Expected Reversals, Unbinned



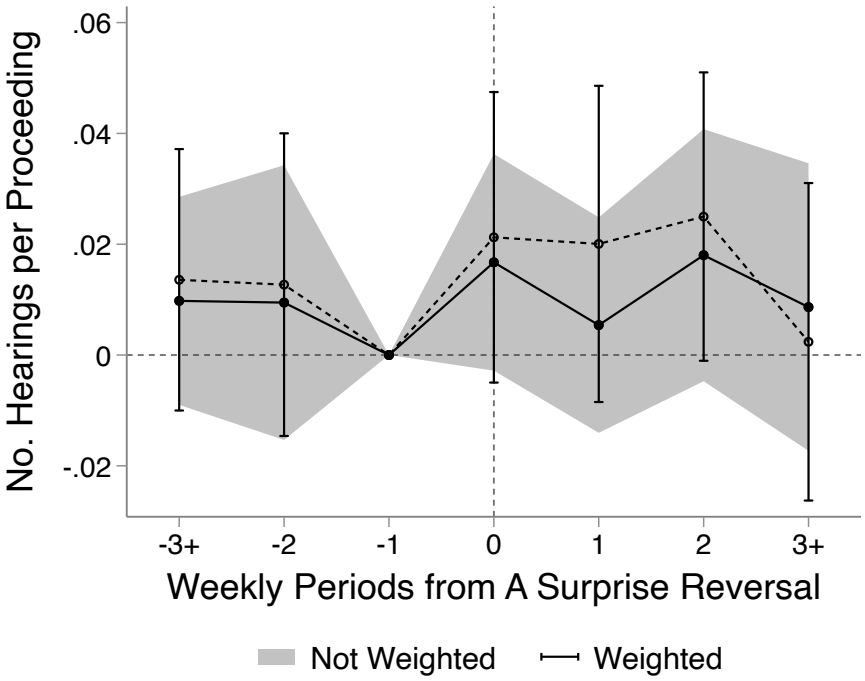
Notes: The figure plots $\hat{b}_k$ from Specification (4.2) using the continuous early-predictability score defined in Section 4.3 as the outcome. Coefficients are estimated separately for each seven-day period $k \in \{-6, \ldots, 6\}$, with period $k = -1$ omitted. Each coefficient measures how the change in early predictability around surprise reversals differs from the corresponding change around expected reversals, relative to the omitted period. The feedback-diagnosticity event sample contains realized reversals with valid classifier assignments: an expected reversal is one for which the classifier predicts an appeal grant, and a surprise reversal is one for which the classifier predicts an appeal denial. The estimation further requires a valid early-predictability score for the linked lower-court proceeding. Lower-court proceedings are linked to events using nearest-event attribution. Period $k = 0$ begins on the recorded BIA decision date and ends six days later. All specifications include BIA event fixed effects, BIA decision-year-by-day-of-week fixed effects, and BIA decision-month-by-day-of-week fixed effects. Standard errors are clustered by immigration judge. Solid dots with shaded pointwise 95% confidence intervals report the unweighted estimates; open circles with capped pointwise 95% confidence intervals report the event-equal-weighted estimates. Section 4.1 describes the linkage and weighting procedures.

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

Measure	Mean	Median	N
Panel A: Proceedings associated with cases in the main analysis			
Hearing records per proceeding	3.79	3	1,227,280
Average adjournment duration (min.)	139.18	122	1,079,585
Share of preparation adjournments	0.08	0	1,103,956
Share of individual-calendar hearings	0.35	0.33	1,227,053
Panel B: All proceedings in the hearing records			
Hearing records per proceeding	2.75	2	8,604,714
Average adjournment duration (min.)	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 proceeding-level summary statistics for the four procedural hearing measures defined in Section 4.3. Hearing records per proceeding counts the records associated with each proceeding. Average adjournment duration is measured in minutes across records with valid start and stop times. The preparation-adjournment share is calculated among records with a nonmissing adjournment-reason classification, and the individual-calendar share is calculated among records with a nonmissing calendar-type classification. Panel A includes proceedings associated with cases in the main analysis. Panel B includes all proceedings observed in the hearing records from 1984 through 2019. Before proceeding-level aggregation, recorded adjournment duration is winsorized at the upper 1% of the full hearing-record distribution. For each row, N is the number of proceedings with a valid value for the indicated measure.

Figure A.11: Hearing Records per Proceeding: Surprise versus Expected Reversals



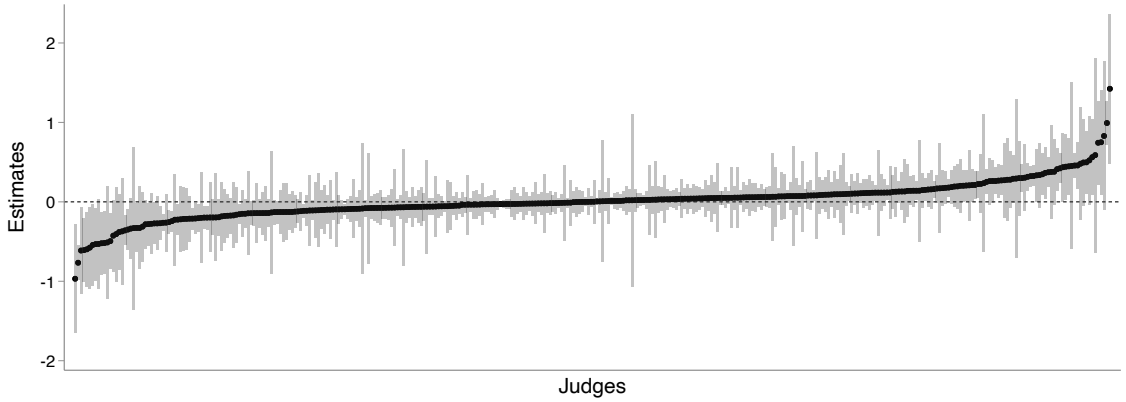
Notes: The figure plots binned estimates of $\hat{\delta}_\kappa$ from Specification (4.3), using the judge-day average number of hearing records per proceeding as the outcome. Each coefficient measures how the change in this outcome around surprise reversals differs from the corresponding change around expected reversals, relative to the omitted period $k = -1$. The feedback-diagnosticity event sample contains realized reversals with valid classifier assignments and a valid judge-day hearing measure. Judge-day observations constructed from proceedings associated with cases in the main analysis are linked to events using nearest-event attribution. The x-axis indexes the binned event-time periods around the recorded BIA decision date; period $k = 0$ begins on that date and ends six days later. Event-equal weights are calculated as described in Section 4.1. All specifications include BIA event fixed effects, BIA decision-year-by-day-of-week fixed effects, and BIA decision-month-by-day-of-week fixed effects. Standard errors are clustered by immigration judge. Solid dots with shaded pointwise 95% confidence intervals report the unweighted estimates; open circles with capped pointwise 95% confidence intervals report the event-equal-weighted estimates.

Table A.7: Broader Procedural Responses: Surprise versus Expected Reversals

	(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

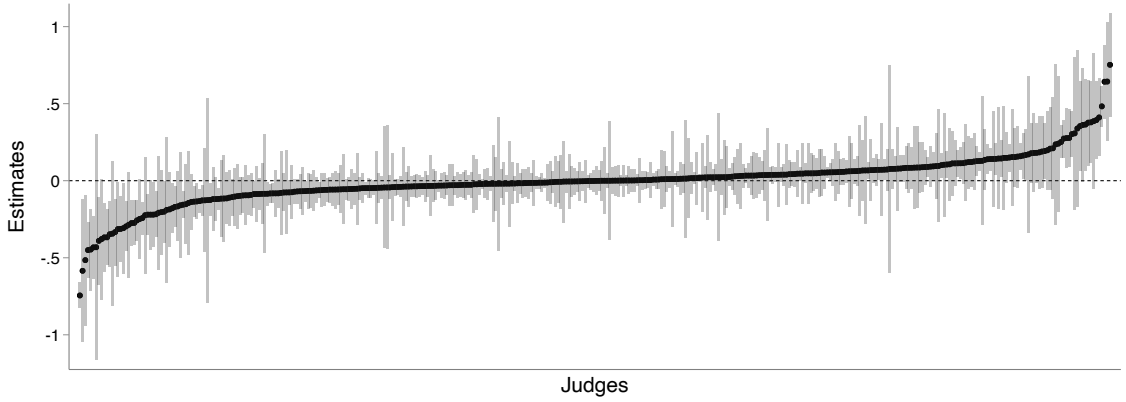
Notes: This table reports unweighted estimates of $\hat{\theta}_\kappa$ and $\hat{\delta}_\kappa$ from Specification (4.3). The four outcomes are judge-day averages constructed from all proceedings observed in the hearing records: hearing records per proceeding, average adjournment duration in minutes, the preparation-adjournment share, and the individual-calendar share. The $\hat{\theta}_\kappa$ rows trace changes around expected reversals relative to the omitted period $k = -1$. The $\hat{\delta}_\kappa$ rows measure how the corresponding changes around surprise reversals differ from those around expected reversals, relative to the same omitted period. The estimation sample links these broader judge-day measures to feedback-diagnosticity events from 1986 through 2019 using nearest-event attribution. All specifications include BIA event fixed effects, BIA decision-year-by-day-of-week fixed effects, and BIA decision-month-by-day-of-week fixed effects. Standard errors clustered by immigration judge are reported in parentheses. The number of judge-day-event-period observations differs across columns because the underlying hearing fields have different rates of valid reporting. Significance levels: 1% (***), 5% (**), and 10% (*).

Figure A.12: Grant-Rate Responsiveness of Immigration Judges



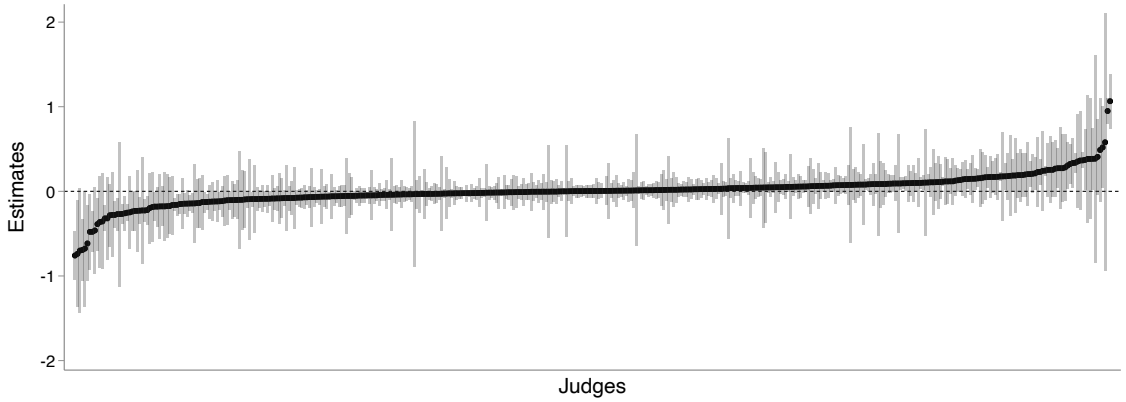
Notes: The figure plots the judge-specific coefficients $\hat{\mu}_{2j}$ estimated from Specification (5.1). The coefficient measures the differential change in grant rates after surprise rather than expected reversals during the pooled post-event window $k \in \{0, 1\}$, relative to the omitted pre-event periods. This window covers the recorded BIA decision date through thirteen days afterward. The outcome is an indicator equal to one if the immigration judge grants the proceeding. The feedback-diagnosticity event sample contains realized reversals from 1986 to 2019 with valid classifier assignments and is linked to lower-court proceedings using the nearest-event-attribution rule described in Section 4.1. The specifications are unweighted and include BIA event fixed effects and the BIA decision calendar fixed effects defined in Section 4.1. Standard errors are robust to heteroskedasticity. Each dot represents one of 354 immigration judges, ranked by the estimated response; gray lines denote 95% confidence intervals.

Figure A.13: Early-Predictability Responsiveness of Immigration Judges



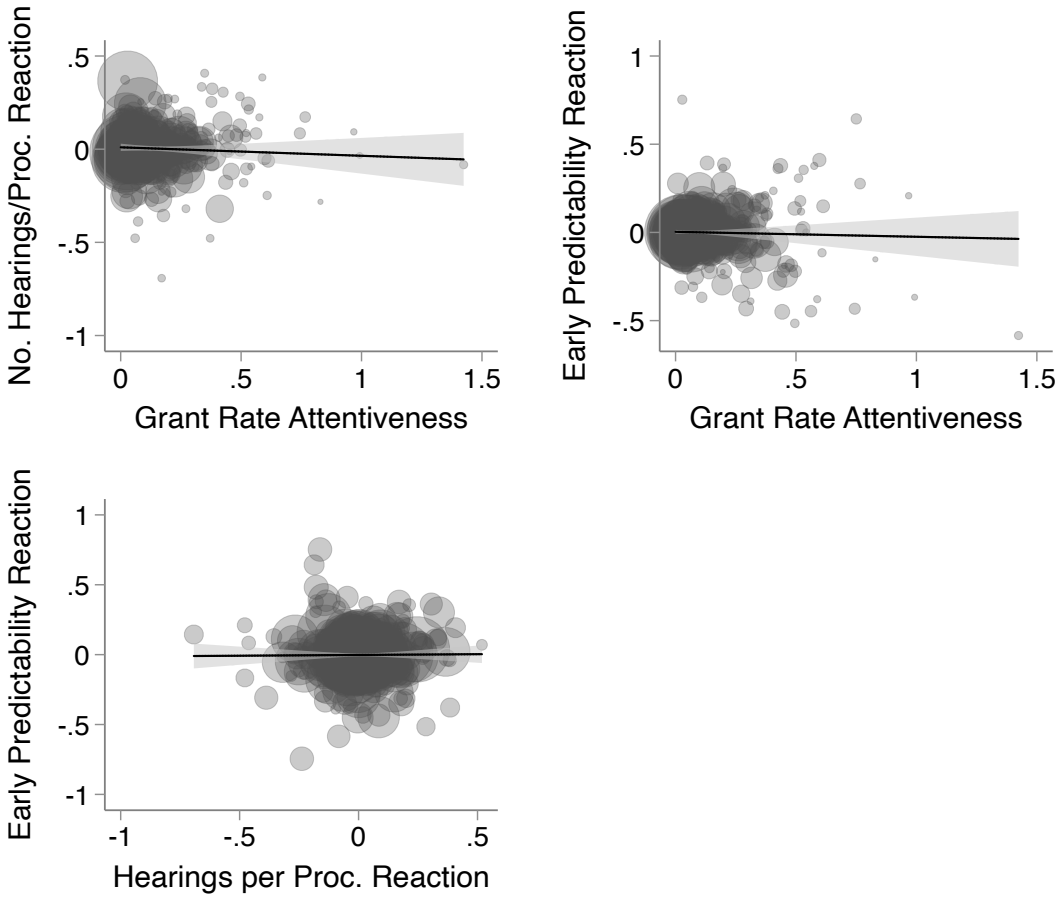
Notes: The figure plots the judge-specific coefficients $\hat{\mu}_{2j}$ estimated from Specification (5.1). The coefficient measures the differential change in early predictability after surprise rather than expected reversals during the pooled post-event window $k \in \{0, 1\}$, relative to the omitted pre-event periods. This window covers the recorded BIA decision date through thirteen days afterward. The outcome is the continuous early-predictability score described in Section 4.3. The feedback-diagnosticity event sample contains realized reversals from 1986 to 2019 with valid classifier assignments and is linked to lower-court proceedings using the nearest-event-attribution rule described in Section 4.1. The specifications are unweighted and include BIA event fixed effects and the BIA decision calendar fixed effects defined in Section 4.1. Standard errors are robust to heteroskedasticity. Each dot represents one of 380 immigration judges, ranked by the estimated response; gray lines denote 95% confidence intervals.

Figure A.14: Hearing-Count Responsiveness of Immigration Judges



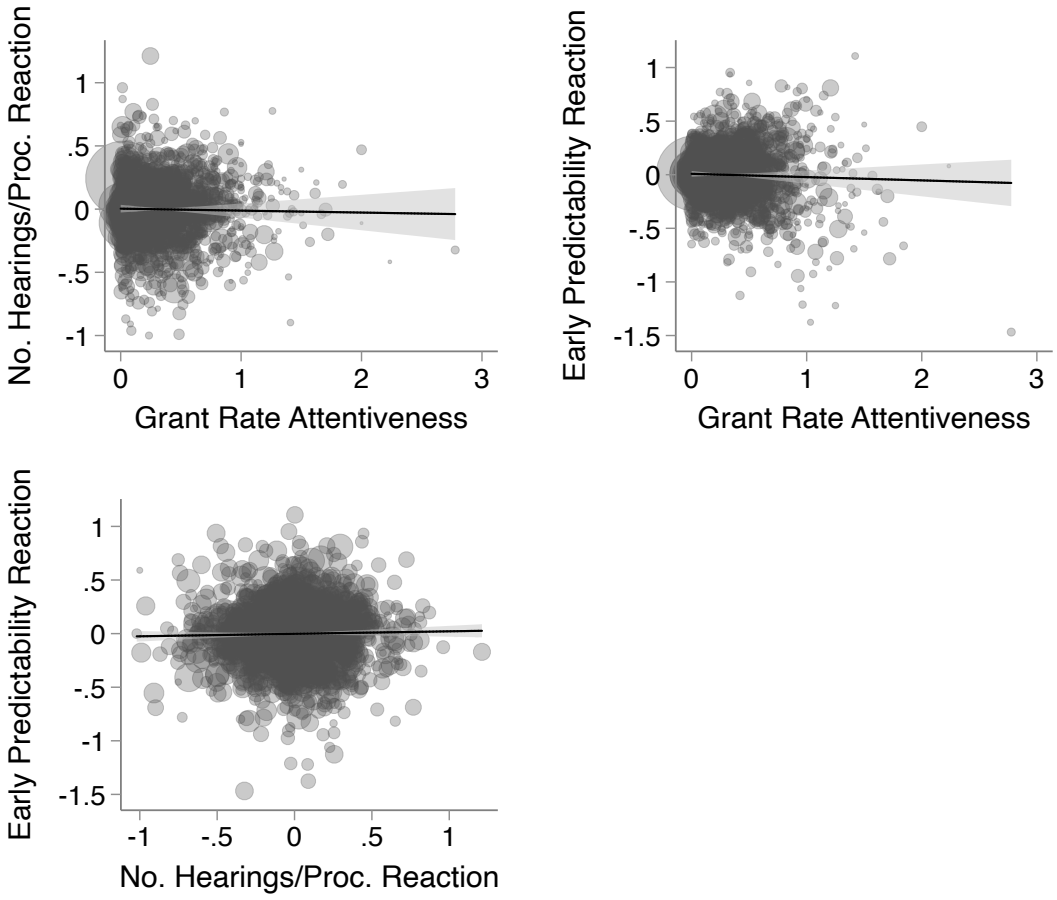
Notes: The figure plots the judge-specific coefficients $\hat{\mu}_{2j}$ estimated from Specification (5.2). The coefficient measures the differential change in hearing counts after surprise rather than expected reversals during the pooled post-event window $k \in \{0, 1\}$, relative to the omitted pre-event periods. This window covers the recorded BIA decision date through thirteen days afterward. The outcome is the daily average number of hearing records per proceeding described in Section 4.3. The feedback-diagnosticity event sample contains realized reversals from 1986 to 2019 with valid classifier assignments and a valid judge-day hearing measure. Judge-day observations are linked to events using the nearest-event-attribution rule described in Section 4.1. The specifications are unweighted and include BIA event fixed effects and the BIA decision calendar fixed effects defined in Section 4.1. Standard errors are robust to heteroskedasticity. Each dot represents one of 416 immigration judges, ranked by the estimated response; gray lines denote 95% confidence intervals.

Figure A.15: Correlations among Judge-Level Responsiveness Measures



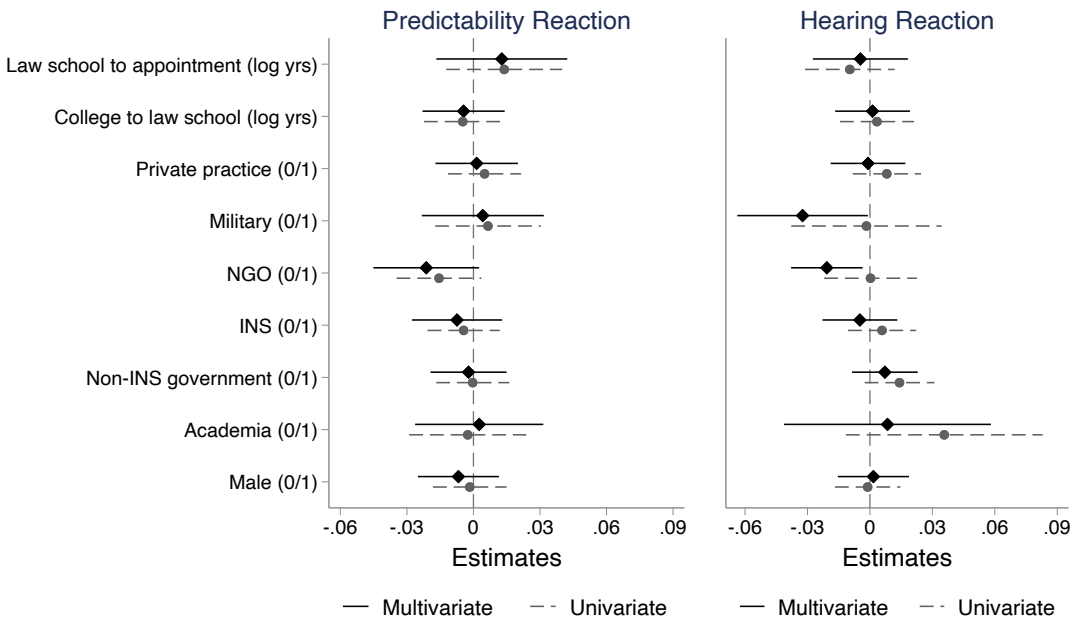
Notes: The figure plots pairwise relationships among the grant-rate, early-predictability, and hearing-count responsiveness measures described in Section 5.2.1. Each dot represents one immigration judge for whom both measures shown in that panel are estimable. The size of each dot reflects the number of observations used to estimate the measure on the horizontal axis. The fitted line is weighted by that number, and the gray shaded area denotes its 95% confidence interval based on heteroskedasticity-robust standard errors.

Figure A.16: Correlations among Judge-Year Responsiveness Measures



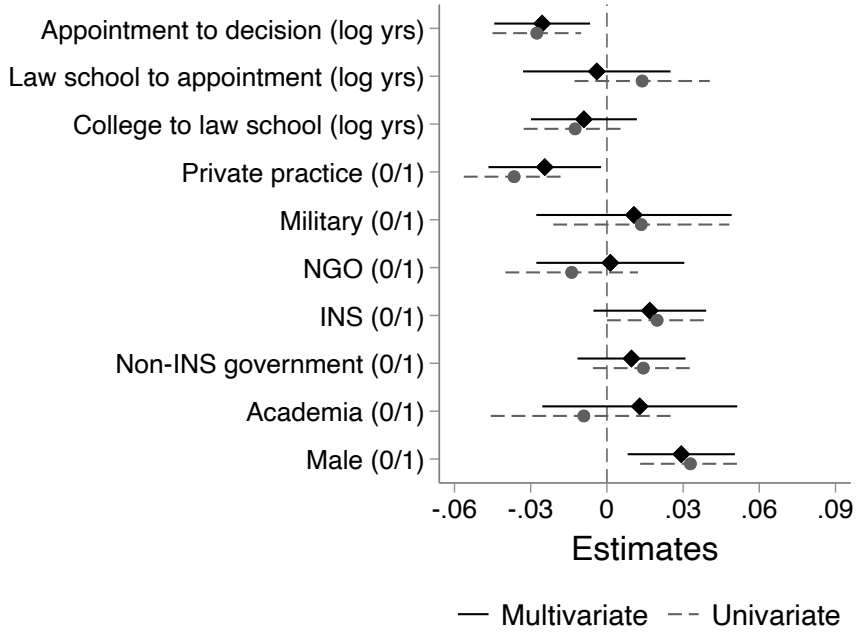
Notes: The figure plots pairwise relationships among the grant-rate, early-predictability, and hearing-count responsiveness measures constructed at the judge-year level. Each dot represents an immigration-judge-year observation for which both measures shown in that panel are estimable. The size of each dot reflects the number of observations used to estimate the measure on the horizontal axis. The fitted line is weighted by that number and includes BIA decision-year fixed effects. The gray shaded area denotes the 95% confidence interval based on heteroskedasticity-robust standard errors.

Figure A.17: Hearing-Process Responsiveness and Judge Characteristics



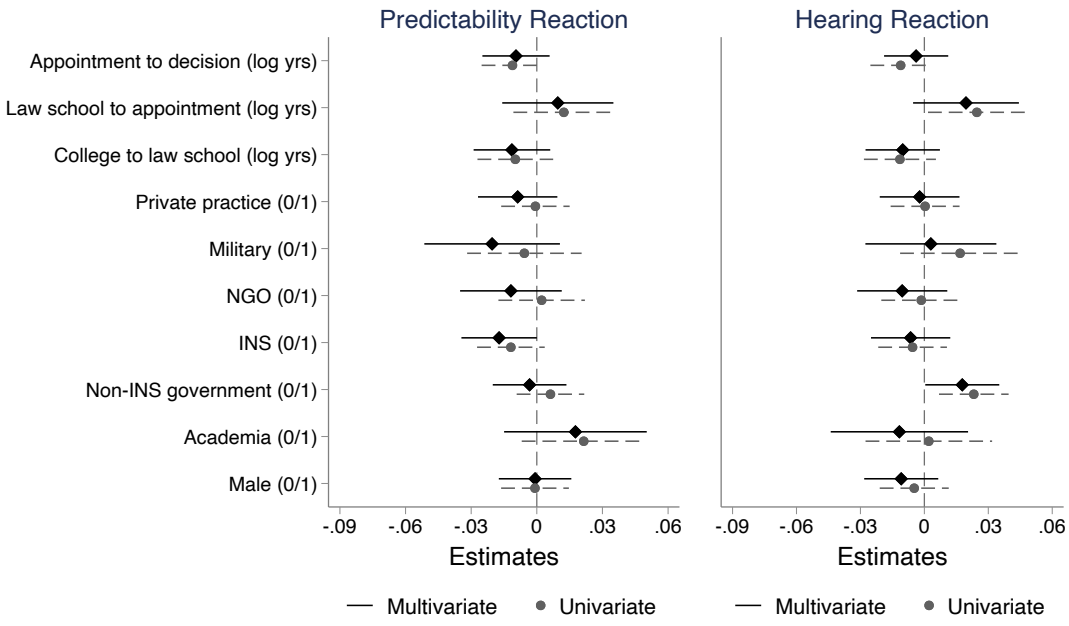
Notes: The figure reports associations between immigration judges' biographical characteristics and the two secondary judge-level responsiveness measures. The left panel reports associations with early-predictability responsiveness, and the right panel reports associations with hearing-count responsiveness. Diamonds represent coefficients from multivariate regressions including all characteristics as regressors, with solid lines denoting 90% confidence intervals. Circles represent coefficients from univariate regressions, with dashed lines denoting 90% confidence intervals. All regressions are weighted by the number of proceedings ruled by each judge. Standard errors are robust to heteroskedasticity.

Figure A.18: Judge-Year Grant-Rate Responsiveness and Characteristics



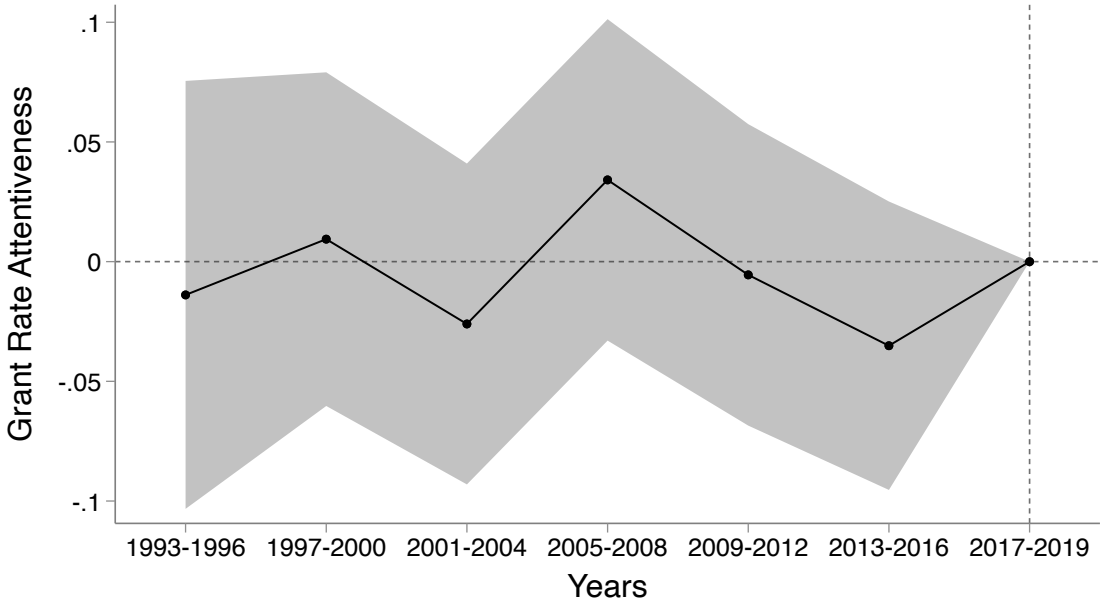
Notes: The figure reports associations between the judge-year grant-rate responsiveness measure and immigration judges' biographical characteristics. Diamonds represent coefficients from multivariate regressions including all characteristics as regressors, with solid lines denoting 90% confidence intervals. Circles represent coefficients from univariate regressions, with dashed lines denoting 90% confidence intervals. All regressions include BIA decision-year fixed effects and are weighted by the number of proceedings ruled by each judge in that year. Standard errors are robust to heteroskedasticity.

Figure A.19: Judge-Year Hearing-Process Responsiveness and Characteristics



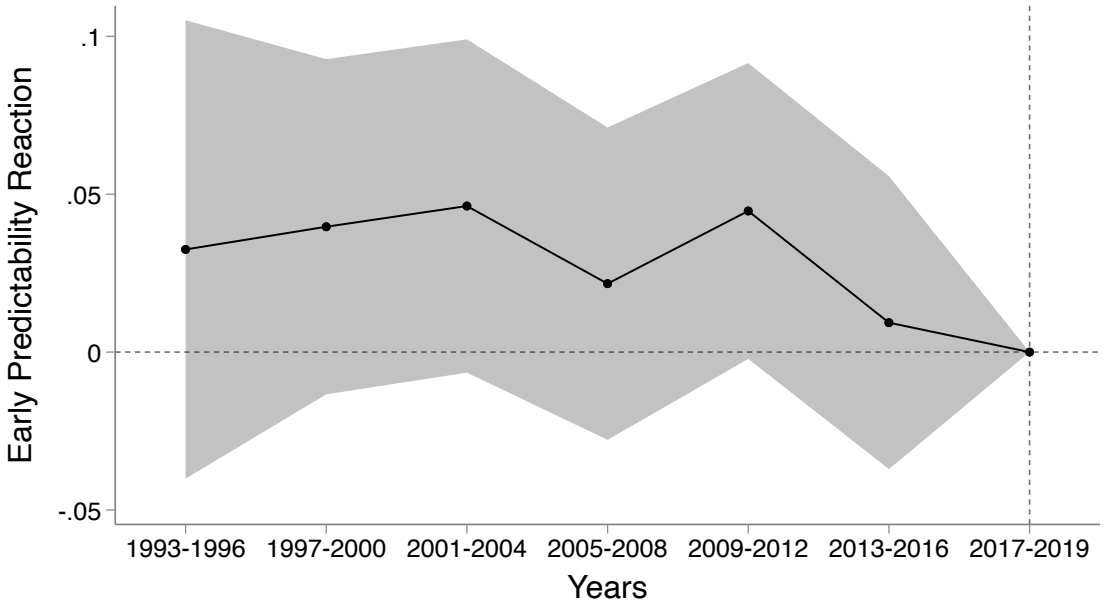
Notes: The figure reports associations between immigration judges' biographical characteristics and the two secondary judge-year responsiveness measures. The left panel reports associations with early-predictability responsiveness, and the right panel reports associations with hearing-count responsiveness. Diamonds represent coefficients from multivariate regressions including all characteristics as regressors, with solid lines denoting 90% confidence intervals. Circles represent coefficients from univariate regressions, with dashed lines denoting 90% confidence intervals. All regressions include BIA decision-year fixed effects and are weighted by the number of proceedings ruled by each judge in that year. Standard errors are robust to heteroskedasticity.

Figure A.20: Grant-Rate Responsiveness across Presidential Periods



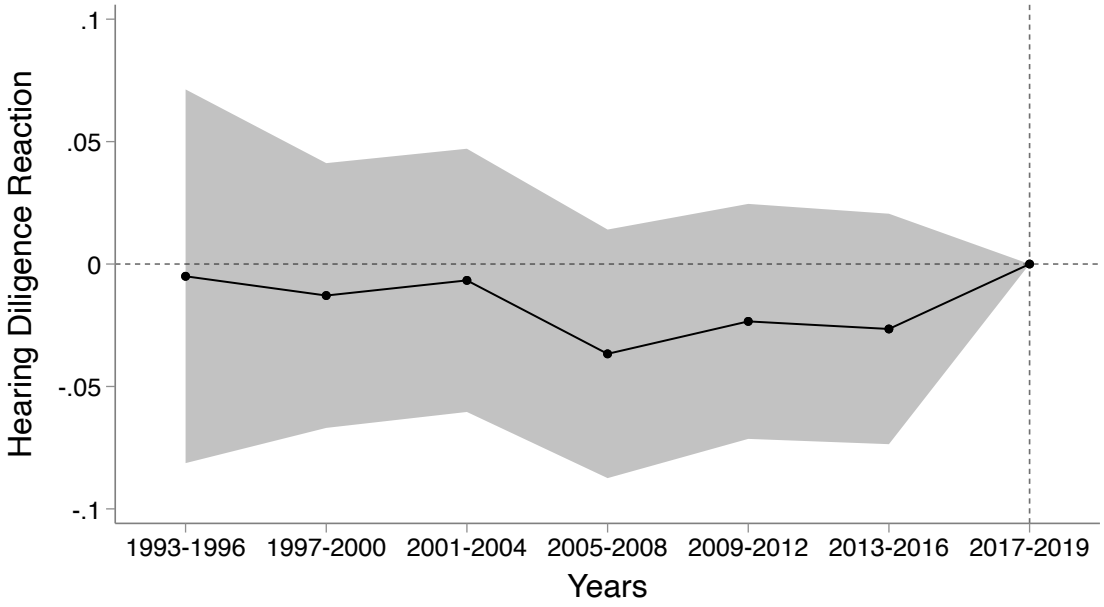
Notes: The figure reports within-judge differences in judge-year grant-rate response magnitude across presidential periods from 1993 to 2019, with 2017 to 2019 as the omitted reference period. Circles represent the estimated period coefficients, and the shaded area denotes 95% confidence intervals. The regression includes immigration-judge fixed effects and weights each judge-year by the number of proceedings ruled by that judge in that year. Standard errors are robust to heteroskedasticity.

Figure A.21: Early-Predictability Responsiveness across Presidential Periods



Notes: The figure reports within-judge differences in signed judge-year early-predictability responses across presidential periods from 1993 to 2019, with 2017 to 2019 as the omitted reference period. Circles represent the estimated period coefficients, and the shaded area denotes 95% confidence intervals. The regression includes immigration-judge fixed effects and weights each judge-year by the number of proceedings ruled by that judge in that year. Standard errors are robust to heteroskedasticity.

Figure A.22: Hearing-Count Responsiveness across Presidential Periods



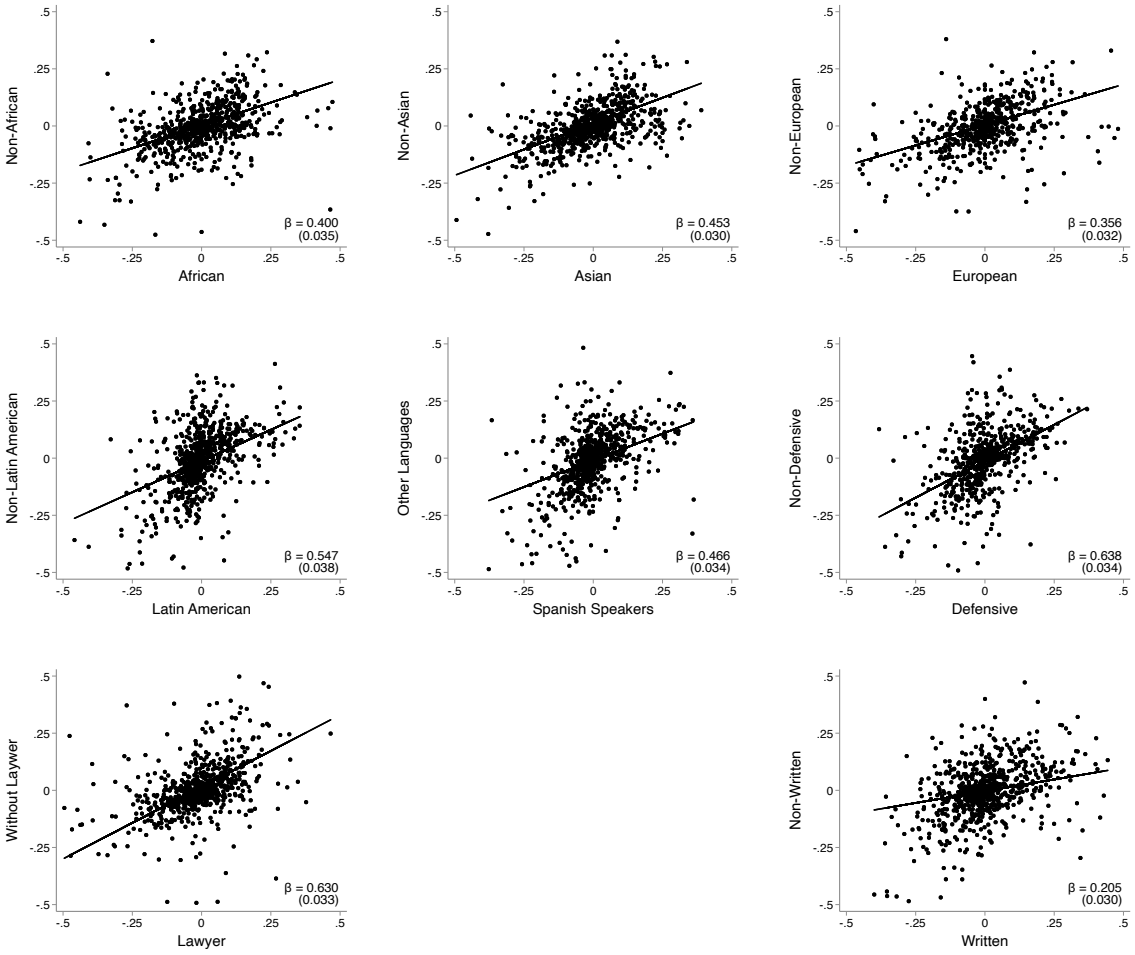
Notes: The figure reports within-judge differences in signed judge-year hearing-count responses across presidential periods from 1993 to 2019, with 2017 to 2019 as the omitted reference period. Circles represent the estimated period coefficients, and the shaded area denotes 95% confidence intervals. The regression includes immigration-judge fixed effects and weights each judge-year by the number of proceedings ruled by that judge in that 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 proceeding	-0.098***	-0.019***
	(0.005)	(0.003)
Written decision	-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 hearing day	0.013***	0.001
	(0.001)	(0.001)
Hearing start hour	-0.005	-0.005
	(0.015)	(0.013)
Hearing start time	0.000	0.000
	(0.000)	(0.000)
Morning hearing	-0.005	-0.014
	(0.018)	(0.018)
Lunchtime hearing	-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 from filing to case opening	0.000***	-0.000
	(0.000)	(0.000)
Joint F-test	239.54	10.21
	[0.000]	[0.000]

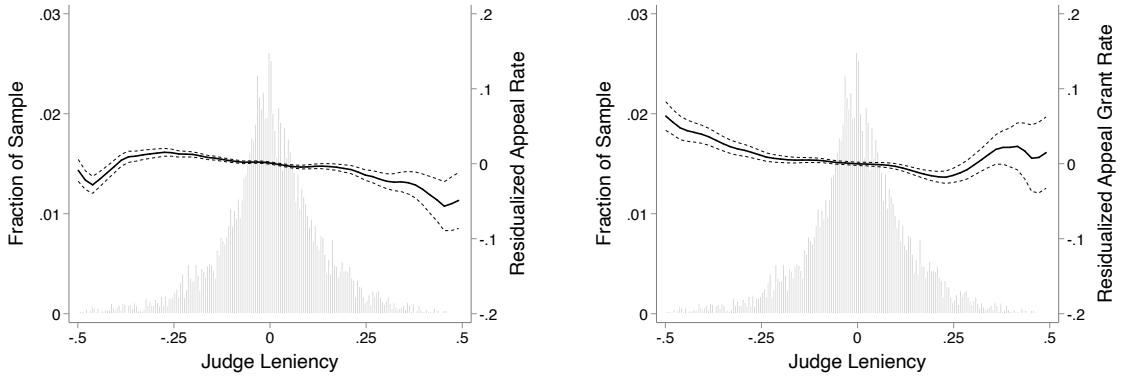
Notes: This table tests whether observed proceeding characteristics predict lower-court grants or the leave-out judge-leniency measure after conditioning on the court-by-time assignment cells. The underlying sample consists of lower-court proceedings decided between 1986 and 2019. Column (1) reports coefficients from an OLS regression of the lower-court grant indicator on the listed characteristics. Column (2) reports coefficients from an OLS regression of the leave-out judge-leniency measure on the same characteristics. Both specifications absorb court-by-decision-year-by-day-of-week and court-by-month-by-day-of-week fixed effects. The bottom rows report the F-statistics and p-values for the joint hypothesis that the coefficients on all listed characteristics equal zero. Standard errors clustered by immigration judge are reported in parentheses. Asterisks denote significance at the 1% (***), 5% (**), and 10% (*) levels.

Figure A.23: Rank Stability of Judge Leniency across Proceeding Groups



Notes: This figure assesses rank stability by comparing leave-out judge-leniency measures constructed separately for the two proceeding groups indicated on the axes of each panel. Each dot represents an immigration judge for whom both group-specific measures are estimable. The horizontal axis reports leniency for the group named on that axis, and the vertical axis reports leniency for the corresponding comparison group. The measures are constructed from lower-court grant indicators residualized using the court-by-time fixed effects described in Section 5.2.2. The solid line reports the fitted relationship from an OLS regression of the vertical-axis measure on the horizontal-axis measure, with the same court-by-time fixed effects. Each panel reports the estimated slope $\hat{\beta}$, with its immigration-judge-clustered standard error in parentheses.

Figure A.24: Judge Leniency and the Two Review Margins

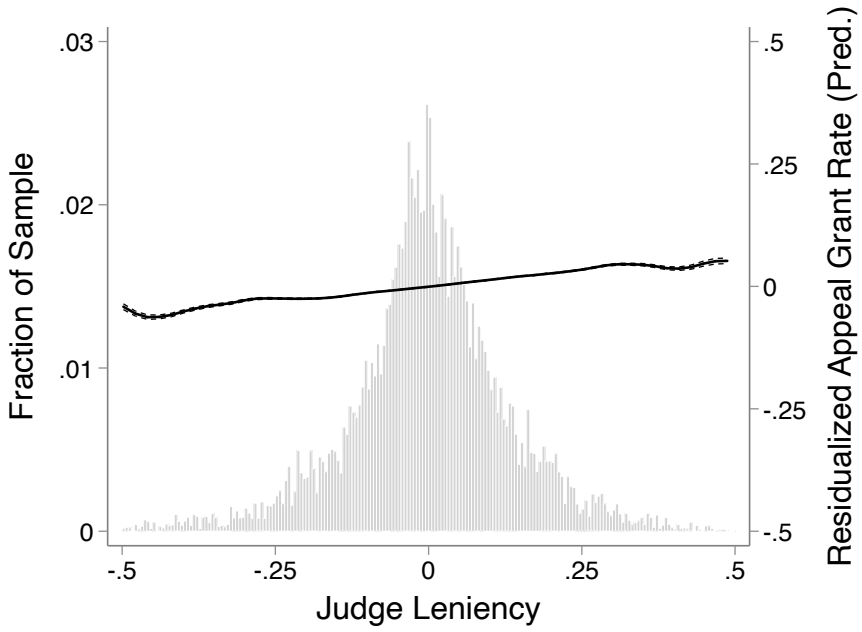


(a) Selection Margin

(b) Success Margin

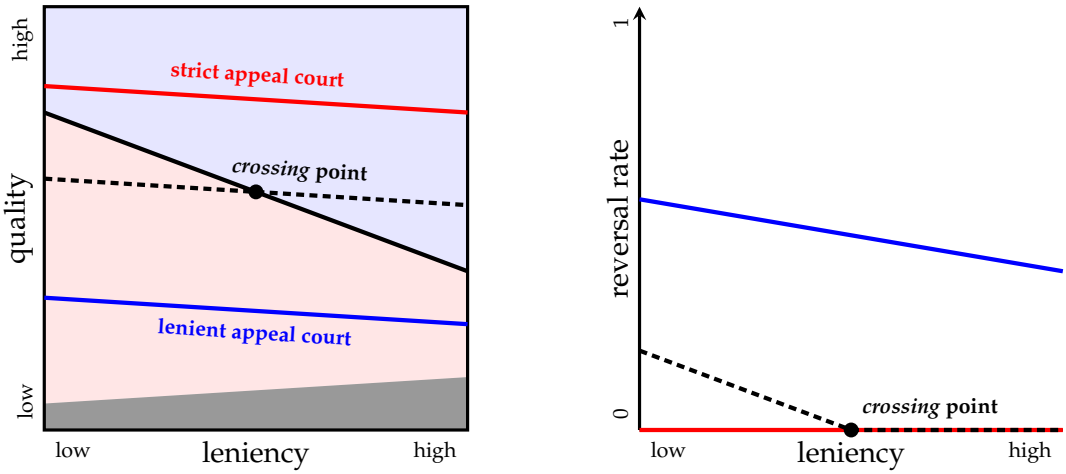
Notes: This figure plots the reduced-form relationships between judge leniency and the two margins through which lower-court denials enter and proceed through appellate review. Panel A.24a reports the selection margin using the lower-court denial sample; the outcome equals one if the denial is appealed and zero otherwise. Panel A.24b reports the success margin using appealed lower-court denials with a BIA disposition classified as an appeal grant or appeal denial; the outcome equals one for an appeal grant and zero for an appeal denial. Neutral or other BIA dispositions are excluded from the success-margin sample. The leave-out judge-leniency measure is constructed as described in Section 5.2.2. Each outcome is residualized using the court-by-time fixed effects used to construct judge leniency. The solid lines report local-linear estimates with bandwidth 0.04, and the dashed lines denote pointwise 95% confidence intervals. The gray bars show the distribution of judge leniency across lower-court proceedings.

Figure A.25: Judge Leniency and Predicted BIA Grant Probability



Notes: This figure plots the relationship between judge leniency and the predicted BIA grant probability generated by the random-forest classifier described in Section 3.2. The estimation sample consists of lower-court proceedings with a valid predicted probability and leave-out judge-leniency measure. The classifier is trained on appeal records with BIA dispositions classified as appeal grants or appeal denials and is then applied to lower-court proceedings, including those without a realized BIA disposition. For proceedings without an observed appeal-filing date, the prediction procedure imputes appeal-filing timing using the median interval between the lower-court decision and appeal filing among observed appeals. The leave-out judge-leniency measure is constructed as described in Section 5.2.2. The predicted BIA grant probability is residualized using the court-by-time fixed effects used to construct judge leniency. The solid line reports a local-linear estimate with bandwidth 0.04, and the dashed lines denote pointwise 95% confidence intervals. The gray bars show the distribution of judge leniency across lower-court proceedings.

Figure A.26: A Stylized Extension with Leniency-Dependent BIA Risk

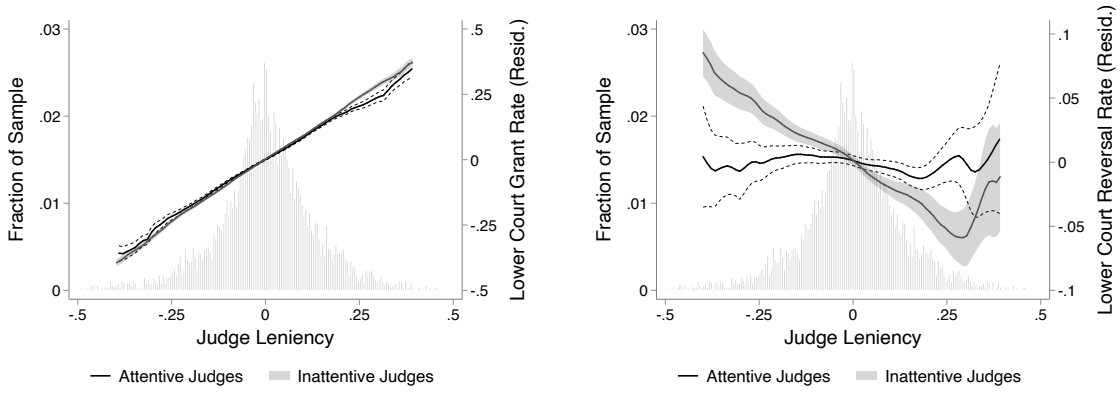


(a) Decision Thresholds

(b) Unconditional Reversal Rates

Notes: This figure presents a stylized extension of the threshold model in Section 5.1 that allows effective BIA grant risk to vary modestly with judge leniency, reflecting the case-composition relationship documented in Appendix Figure A.25. In Panel A.26a, the solid black line represents the immigration-judge decision threshold. The red, blue, and dashed black lines illustrate strict, lenient, and crossing cases for the effective BIA threshold. The blue region contains first-instance grants, the red region contains denials above the appeal-filing threshold, and the gray region contains denials below that threshold that do not enter the applicant-appeal channel. Panel A.26b plots the corresponding unconditional reversal rate $R(c)$ in each case. The baseline model retains the common BIA standard c^*, and all coordinates and slopes in this extension are schematic rather than estimated.

Figure A.27: Judge Leniency and Lower-Court Outcomes by Judge-Year Grant-Rate Attentiveness

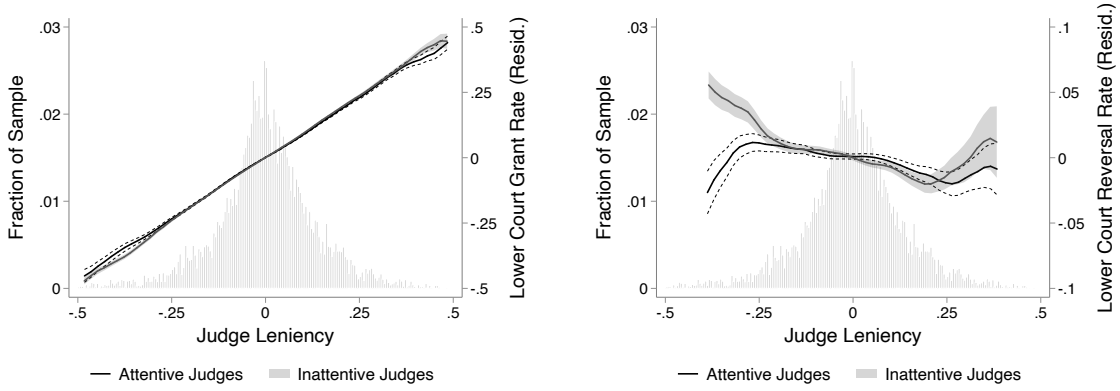


(a) Lower-Court Grant Rate

(b) Reversal Rate among Denials

Notes: The figure relates judge leniency to lower-court outcomes separately for proceedings classified as attentive or inattentive using judge-year grant-rate response estimates. Attentiveness is measured by the magnitude of the judge-year grant-rate response to surprise reversals; judge-year observations in the top tercile are classified as attentive, while those in the bottom tercile are classified as inattentive. Panel A.27a shows local-linear fits of the residualized lower-court grant indicator on judge leniency. Panel A.27b shows the corresponding fits for the reversal indicator among lower-court denials. Both outcomes are residualized using the same court-by-time fixed effects used to construct the leave-out judge-leniency measure described in Section 5.2.2. The gray bars show the distribution of judge leniency across proceedings. The black line represents attentive judges, and the gray line represents inattentive judges; dashed lines and shaded areas denote pointwise 95% confidence intervals. Standard errors are clustered by immigration judge.

Figure A.28: Judge Leniency and Lower-Court Outcomes by Grant-Rate Attentiveness: Top and Bottom Halves

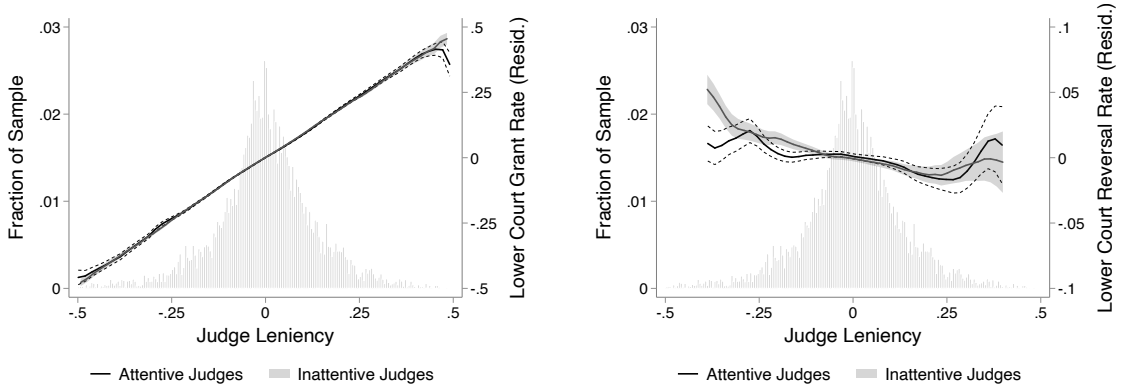


(a) Lower-Court Grant Rate

(b) Reversal Rate among Denials

Notes: The figure relates judge leniency to lower-court outcomes using an alternative top-and-bottom-half classification of grant-rate attentiveness. Attentiveness is measured by the magnitude of each judge's grant-rate response to surprise reversals. Within each ruling year, proceedings in the top half of grant-rate response magnitude are classified as attentive, while those in the bottom half are classified as inattentive. Panel A.28a shows local-linear fits of the residualized lower-court grant indicator on judge leniency. Panel A.28b shows the corresponding fits for the reversal indicator among lower-court denials. Both outcomes are residualized using the same court-by-time fixed effects used to construct the leave-out judge-leniency measure described in Section 5.2.2. The gray bars show the distribution of judge leniency across proceedings. The black line represents attentive judges, and the gray line represents inattentive judges; dashed lines and shaded areas denote pointwise 95% confidence intervals. Standard errors are clustered by immigration judge.

Figure A.29: Judge Leniency and Lower-Court Outcomes by Early-Predictability Attentiveness

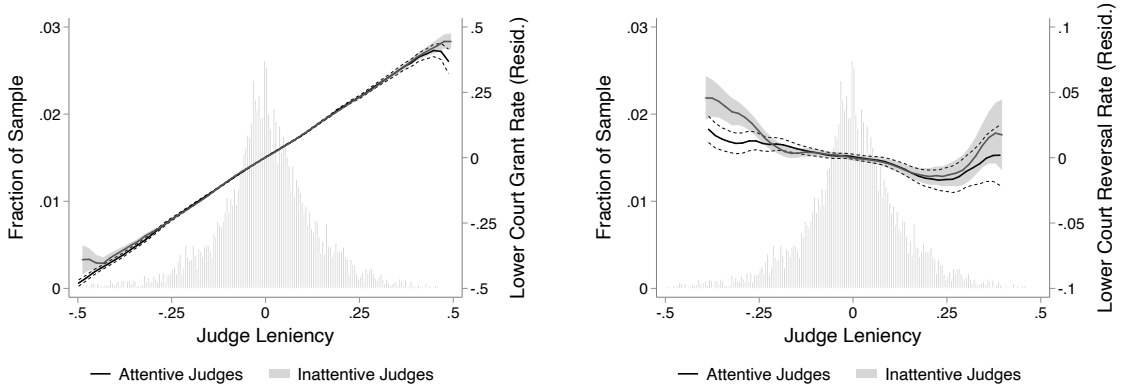


(a) Lower-Court Grant Rate

(b) Reversal Rate among Denials

Notes: The figure relates judge leniency to lower-court outcomes separately for judges whose early predictability decreases or increases after surprise reversals. A decrease in early predictability is classified as greater attentiveness in the hearing process, while an increase is classified as lower attentiveness, as described in Section 5.2.1. Panel A.29a shows local-linear fits of the residualized lower-court grant indicator on judge leniency. Panel A.29b shows the corresponding fits for the reversal indicator among lower-court denials. Both outcomes are residualized using the same court-by-time fixed effects used to construct the leave-out judge-leniency measure described in Section 5.2.2. The gray bars show the distribution of judge leniency across proceedings. The black line represents attentive judges, and the gray line represents inattentive judges; dashed lines and shaded areas denote pointwise 95% confidence intervals. Standard errors are clustered by immigration judge.

Figure A.30: Judge Leniency and Lower-Court Outcomes by Hearing-Count Attentiveness

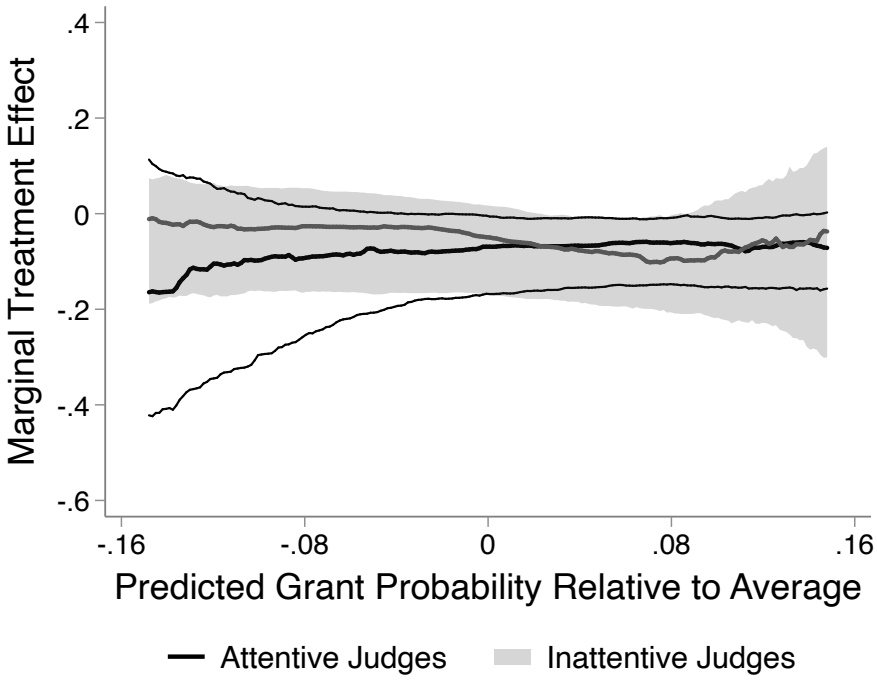


(a) Lower-Court Grant Rate

(b) Reversal Rate among Denials

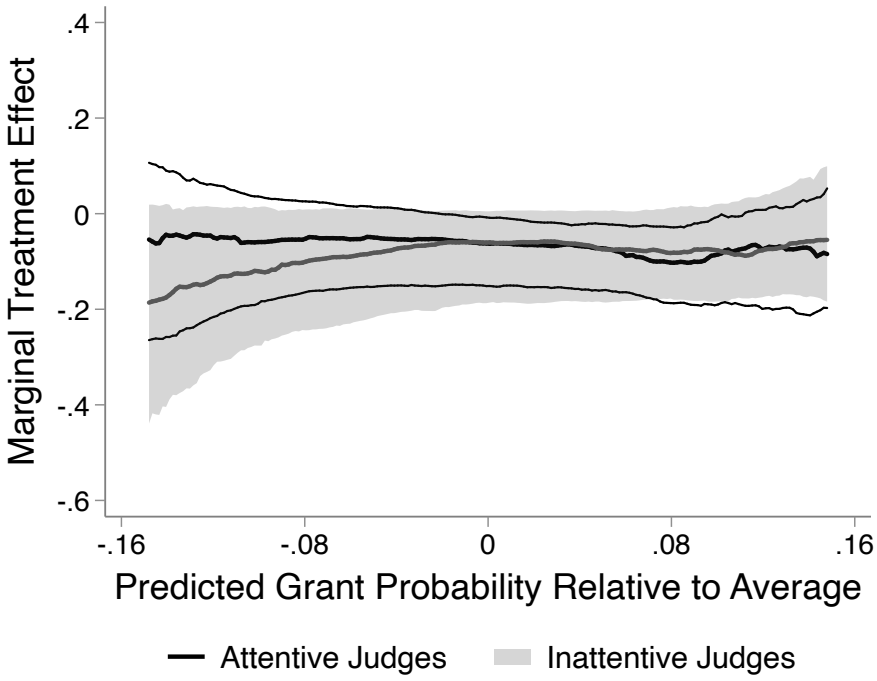
Notes: The figure relates judge leniency to lower-court outcomes separately for judges whose number of hearings per proceeding increases or decreases after surprise reversals. An increase in hearings per proceeding is classified as greater attentiveness in the hearing process, while a decrease is classified as lower attentiveness, as described in Section 5.2.1. Panel A.30a shows local-linear fits of the residualized lower-court grant indicator on judge leniency. Panel A.30b shows the corresponding fits for the reversal indicator among lower-court denials. Both outcomes are residualized using the same court-by-time fixed effects used to construct the leave-out judge-leniency measure described in Section 5.2.2. The gray bars show the distribution of judge leniency across proceedings. The black line represents attentive judges, and the gray line represents inattentive judges; dashed lines and shaded areas denote pointwise 95% confidence intervals. Standard errors are clustered by immigration judge.

Figure A.31: Marginal Treatment Effects by Early-Predictability Attentiveness



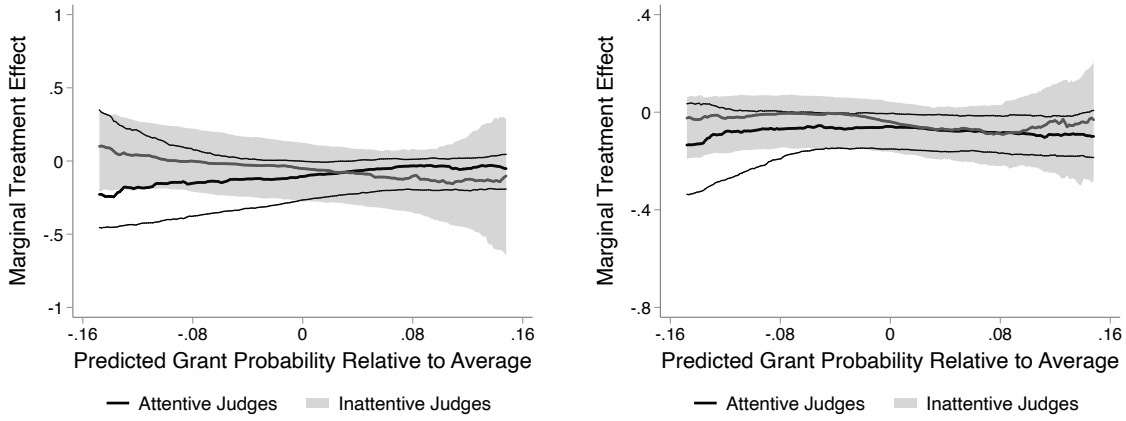
Notes: The figure plots the marginal treatment effect defined in Equation (5.4), separately for proceedings ruled by judges whose early predictability decreases after surprise reversals, classified as attentive, and judges whose early predictability increases, classified as inattentive. The estimation sample consists of lower-court denials, and the outcome equals one when a denial is followed by an appeal grant and zero otherwise. The MTEs follow the residualized local-linear procedure described in Section 5.3.2, using a bandwidth of 0.08 and the same fixed, mean-centered grid of predicted lower-court grant probabilities for both groups. Lower values on the horizontal axis correspond to lower judge leniency and stricter decision thresholds. The solid black line represents attentive judges, and the solid gray line represents inattentive judges. The dashed black lines and gray shaded area report pointwise 90% bootstrap percentile confidence intervals based on 500 replications resampled at the judge level.

Figure A.32: Marginal Treatment Effects by Hearing-Count Attentiveness



Notes: The figure plots the marginal treatment effect defined in Equation (5.4), separately for proceedings ruled by judges who increase the average number of hearings per proceeding after surprise reversals, classified as attentive, and judges who decrease the average number, classified as inattentive. The estimation sample consists of lower-court denials, and the outcome equals one when a denial is followed by an appeal grant and zero otherwise. The MTEs follow the residualized local-linear procedure described in Section 5.3.2, using a bandwidth of 0.08 and the same fixed, mean-centered grid of predicted lower-court grant probabilities for both groups. Lower values on the horizontal axis correspond to lower judge leniency and stricter decision thresholds. The solid black line represents attentive judges, and the solid gray line represents inattentive judges. The dashed black lines and gray shaded area report pointwise 90% bootstrap percentile confidence intervals based on 500 replications resampled at the judge level.

Figure A.33: Marginal Treatment Effects on Appeal Filing and BIA Success by Early-Predictability Attentiveness

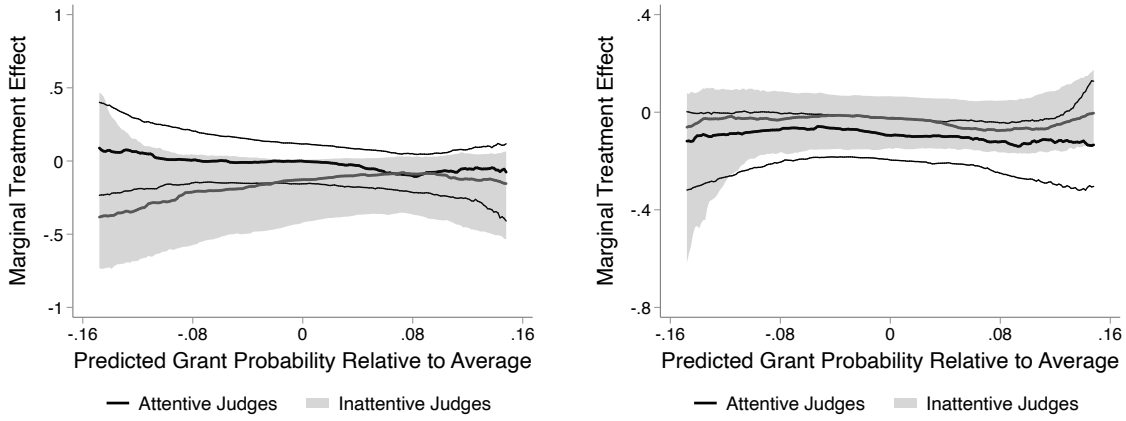


(a) Selection Margin

(b) Success Margin

Notes: The figure plots marginal treatment effects for the two review margins separately for attentive and inattentive judges. Under the early-predictability classification, judges whose early predictability decreases after surprise reversals are classified as attentive, while judges whose early predictability increases are classified as inattentive. Panel A.33a uses lower-court denials and an outcome equal to one if the denial is appealed and zero otherwise. Panel A.33b uses appealed lower-court denials with a BIA disposition classified as an appeal grant or appeal denial; the outcome equals one for an appeal grant and zero for an appeal denial, and neutral or other dispositions are excluded. Both panels apply the residualized local-linear MTE procedure described in Section 5.3.2, using a bandwidth of 0.08 and the same fixed, mean-centered grid of predicted lower-court grant probabilities. The solid black line represents attentive judges, and the solid gray line represents inattentive judges. The dashed black lines and gray shaded areas report pointwise 90% bootstrap percentile confidence intervals based on 500 replications resampled at the judge level.

Figure A.34: Marginal Treatment Effects on Appeal Filing and BIA Success by Hearing-Count Attentiveness



(a) Selection Margin

(b) Success Margin

Notes: The figure plots marginal treatment effects for the two review margins separately for attentive and inattentive judges. Under the hearing-count classification, judges who increase the average number of hearings per proceeding after surprise reversals are classified as attentive, while judges who decrease the average number are classified as inattentive. Panel A.34a uses lower-court denials and an outcome equal to one if the denial is appealed and zero otherwise. Panel A.34b uses appealed lower-court denials with a BIA disposition classified as an appeal grant or appeal denial; the outcome equals one for an appeal grant and zero for an appeal denial, and neutral or other dispositions are excluded. Both panels apply the residualized local-linear MTE procedure described in Section 5.3.2, using a bandwidth of 0.08 and the same fixed, mean-centered grid of predicted lower-court grant probabilities. The solid black line represents attentive judges, and the solid gray line represents inattentive judges. The dashed black lines and gray shaded areas report pointwise 90% bootstrap percentile confidence intervals based on 500 replications resampled at the judge level.

Table A.9: Average Marginal Treatment Effects by Early-Predictability Attentiveness

	(1) Reversal Outcome	(2) Selection Margin	(3) Success Margin
Panel A: Common trimmed support, $p = [-0.1480, 0.1480]$			
$\overline{\text{MTE}}_{\text{att}}$	-0.0808	-0.1036	-0.0783
	[-0.1608, 0.0285]	[-0.2335, 0.0527]	[-0.1323, 0.0073]
$\overline{\text{MTE}}_{\text{inatt}}$	-0.0518	-0.0532	-0.0389
	[-0.1125, 0.0606]	[-0.2425, 0.1544]	[-0.1120, 0.0836]
$\overline{\text{MTE}}_{\text{att}} - \overline{\text{MTE}}_{\text{inatt}}$	-0.0290	-0.0504	-0.0394
	[-0.1635, 0.0949]	[-0.2831, 0.2128]	[-0.1688, 0.0836]
sup-t test	0.9008	1.1968	0.8156
global bootstrap p-value	0.8962	0.6846	0.9281
Panel B: Strict region, $p = [-0.1480, -0.0007]$			
$\overline{\text{MTE}}_{\text{att}}$	-0.0964	-0.1526	-0.0757
	[-0.2151, 0.0594]	[-0.3921, 0.0576]	[-0.1499, 0.0394]
$\overline{\text{MTE}}_{\text{inatt}}$	-0.0286	0.0022	-0.0148
	[-0.0999, 0.1033]	[-1.1961, 0.2045]	[-0.0859, 0.1179]
$\overline{\text{MTE}}_{\text{att}} - \overline{\text{MTE}}_{\text{inatt}}$	-0.0679	-0.1548	-0.0609
	[-0.2390, 0.0916]	[-0.4375, 0.1480]	[-0.2038, 0.0756]
sup-t test	0.9008	1.1968	0.8156
global bootstrap p-value	0.6966	0.4691	0.8184

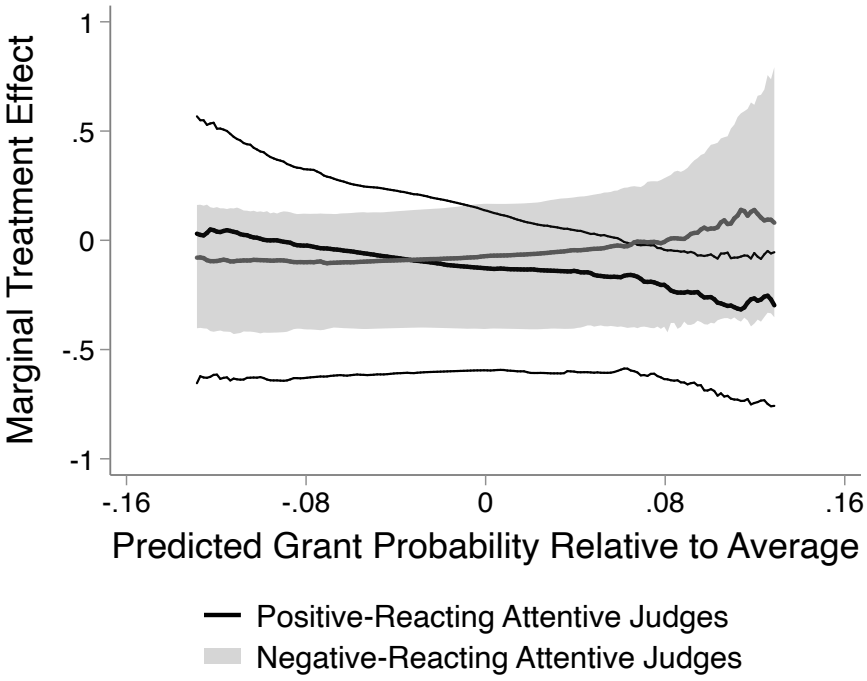
Notes: This table reports uniformly weighted common-support average marginal treatment effects (MTEs) for proceedings ruled by attentive and inattentive judges. Under the early-predictability classification, judges whose early predictability decreases after surprise reversals are classified as attentive, while judges whose early predictability increases are classified as inattentive. For each judge group, the average MTE integrates the estimated MTE function over the indicated mean-centered predicted lower-court grant-probability region and divides by the width of that region. Panel A uses the full common trimmed support, while Panel B uses the lower half of that support, corresponding to the strict region emphasized by the model in Section 5.1. Column (1) uses the denial-conditioned reversal outcome plotted in Figure A.31: the sample consists of lower-court denials, and the outcome equals one when a denial is followed by an appeal grant and zero otherwise. Column (2) uses the selection-margin outcome plotted in Figure A.33a: the sample consists of lower-court denials, and the outcome equals one when the denial is appealed and zero otherwise. Column (3) uses the success-margin outcome plotted in Figure A.33b: the sample consists of appealed lower-court denials classified as appeal grants or appeal denials, and neutral or other dispositions are excluded. Brackets report 90% reverse-percentile confidence intervals based on 500 bootstrap replications clustered at the judge level. The attentive-minus-inattentive row reports the scalar difference between the two average MTEs. The sup-t statistic and global bootstrap p-value test equality of the two MTE functions at every point over the panel-specific support.

Table A.10: Average Marginal Treatment Effects by Hearing-Count Attentiveness

	(1) Reversal Outcome	(2) Selection Margin	(3) Success Margin
Panel A: Common trimmed support, $p = [-0.1480, 0.1480]$			
$\overline{\text{MTE}}_{\text{att}}$	-0.0656	-0.0187	-0.0961
	[-0.1211, 0.0314]	[-0.1464, 0.1341]	[-0.1493, 0.0182]
$\overline{\text{MTE}}_{\text{inatt}}$	-0.0870	-0.1710	-0.0371
	[-0.1660, 0.0280]	[-0.3705, 0.1065]	[-0.1225, 0.0961]
$\overline{\text{MTE}}_{\text{att}} - \overline{\text{MTE}}_{\text{inatt}}$	0.0214	0.1524	-0.0590
	[-0.0937, 0.1439]	[-0.1461, 0.4111]	[-0.1941, 0.0843]
sup-t test	0.8128	1.2813	0.8400
global bootstrap p-value	0.9242	0.6507	0.9082
Panel B: Strict region, $p = [-0.1480, -0.0007]$			
$\overline{\text{MTE}}_{\text{att}}$	-0.0532	0.0172	-0.0794
	[-0.1293, 0.0678]	[-0.1573, 0.1930]	[-0.1437, 0.0526]
$\overline{\text{MTE}}_{\text{inatt}}$	-0.1038	-0.2356	-0.0238
	[-0.2099, 0.0383]	[-0.5522, 0.0812]	[-0.1161, 0.1634]
$\overline{\text{MTE}}_{\text{att}} - \overline{\text{MTE}}_{\text{inatt}}$	0.0506	0.2528	-0.0557
	[-0.1049, 0.2063]	[-0.0946, 0.6231]	[-0.2288, 0.1018]
sup-t test	0.8128	1.2813	0.8266
global bootstrap p-value	0.7206	0.4112	0.6946

Notes: This table reports uniformly weighted common-support average marginal treatment effects (MTEs) for proceedings ruled by attentive and inattentive judges. Under the hearing-count classification, judges who increase the average number of hearings per proceeding after surprise reversals are classified as attentive, while judges who decrease the average number are classified as inattentive. For each judge group, the average MTE integrates the estimated MTE function over the indicated mean-centered predicted lower-court grant-probability region and divides by the width of that region. Panel A uses the full common trimmed support, while Panel B uses the lower half of that support, corresponding to the strict region emphasized by the model in Section 5.1. Column (1) uses the denial-conditioned reversal outcome plotted in Figure A.32: the sample consists of lower-court denials, and the outcome equals one when a denial is followed by an appeal grant and zero otherwise. Column (2) uses the selection-margin outcome plotted in Figure A.34a: the sample consists of lower-court denials, and the outcome equals one when the denial is appealed and zero otherwise. Column (3) uses the success-margin outcome plotted in Figure A.34b: the sample consists of appealed lower-court denials classified as appeal grants or appeal denials, and neutral or other dispositions are excluded. Brackets report 90% reverse-percentile confidence intervals based on 500 bootstrap replications clustered at the judge level. The attentive-minus-inattentive row reports the scalar difference between the two average MTEs. The sup-t statistic and global bootstrap p-value test equality of the two MTE functions at every point over the panel-specific support.

Figure A.35: Marginal Treatment Effects among Attentive Judges by Grant-Rate Response Direction



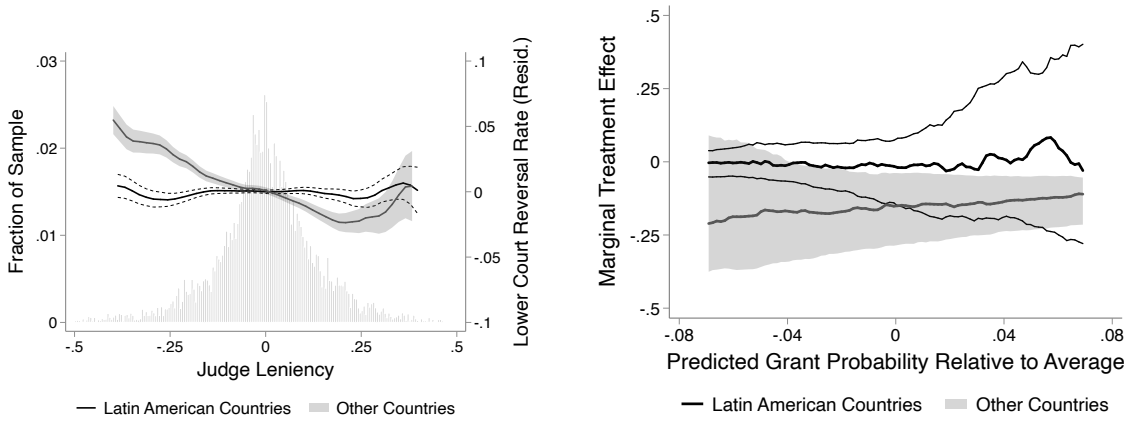
Notes: The figure plots the marginal treatment effect defined in Equation (5.4) for positive-reacting and negative-reacting attentive judges. Attentive judges are identified by the magnitude of their grant-rate response to surprise reversals, as described in Section 5.2.1, and are then separated by the direction of that response. Positive-reacting judges increase their grant rates after surprise reversals, while negative-reacting judges decrease their grant rates. The estimation sample consists of lower-court denials, and the outcome equals one when a denial is followed by an appeal grant and zero otherwise. The MTEs follow the residualized local-linear procedure described in Section 5.3.2, using a bandwidth of 0.08 and the same mean-centered grid of predicted lower-court grant probabilities for both groups. The figure displays the common grid over $|p| \leq 0.13$. The solid black line represents positive-reacting attentive judges, and the solid gray line represents negative-reacting attentive judges. The dashed black lines and gray shaded area report pointwise 90% bootstrap percentile confidence intervals based on 500 replications resampled at the judge level.

Table A.11: Average Marginal Treatment Effects by Procedural Type and Attentiveness

	(1) All	(2) Attentive Judges	(3) Inattentive Judges
Panel A: Common trimmed support, $p = [-0.0796, 0.0796]$			
$\overline{\text{MTE}}_{\text{defensive}}$	-0.0318	0.0205	-0.1025
	[-0.0704, 0.0005]	[-0.1280, 0.2064]	[-0.1983, 0.0002]
$\overline{\text{MTE}}_{\text{affirmative}}$	-0.0990	-0.1920	-0.5451
	[-0.1731, 0.0394]	[-0.3148, 0.0132]	[-0.8928, -0.2518]
$\overline{\text{MTE}}_{\text{defensive}} - \overline{\text{MTE}}_{\text{affirmative}}$	0.0672	0.2125	0.4426
	[-0.0793, 0.1399]	[-0.0405, 0.4558]	[0.1465, 0.8221]
sup-t test	1.6652	1.4607	1.3743
global bootstrap p-value	0.4012	0.4431	0.2954
Panel B: Strict region, $p = [-0.0796, -0.0007]$			
$\overline{\text{MTE}}_{\text{defensive}}$	-0.0405	-0.0132	-0.1006
	[-0.0710, -0.0028]	[-0.1470, 0.1873]	[-0.1883, 0.0096]
$\overline{\text{MTE}}_{\text{affirmative}}$	-0.1447	-0.2247	-0.6473
	[-0.2461, 0.0441]	[-0.3857, -0.0093]	[-1.1549, -0.0743]
$\overline{\text{MTE}}_{\text{defensive}} - \overline{\text{MTE}}_{\text{affirmative}}$	0.1041	0.2115	0.5468
	[-0.0866, 0.2108]	[-0.0401, 0.4915]	[-0.0257, 1.0621]
sup-t test	1.6652	1.4607	1.3299
global bootstrap p-value	0.2156	0.2814	0.1457

Notes: This table reports uniformly weighted average marginal treatment effects (MTEs) for defensive and affirmative proceedings, both for all judges and separately by grant-rate attentiveness. The estimation sample consists of lower-court denials, and the outcome equals one when a denial is followed by an appeal grant and zero otherwise. Column (1) pools all judges, while Columns (2) and (3) report proceedings ruled by attentive and inattentive judges, respectively. Attentive and inattentive judges correspond to the top and bottom terciles of grant-rate response magnitude within ruling year, as described in Section 5.2.1. Each average MTE integrates the estimated MTE function over the indicated mean-centered predicted grant-probability region and divides that integral by the width of the region, following Section 5.3.2. Panel A uses the full common trimmed support for the procedural comparison. Panel B uses the lower half of that support, corresponding to the strict-judge region emphasized by the model in Section 5.1. The difference rows report the average MTE for defensive proceedings minus the average MTE for affirmative proceedings. Brackets report 90% reverse-percentile confidence intervals based on 500 judge-level bootstrap replications. Within each column, the sup-t statistic and global bootstrap p-value test whether the defensive- and affirmative-proceeding MTE functions are equal at every point over the panel-specific support.

Figure A.36: Reversal Rates and MTEs by Latin American Origin

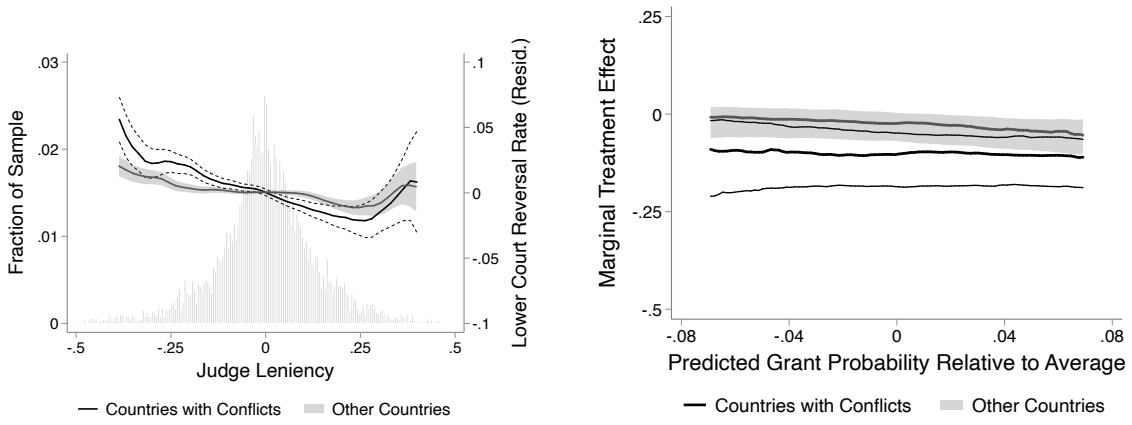


(a) Reduced Form

(b) Marginal Treatment Effect

Notes: This figure compares proceedings involving applicants from Latin American origins with proceedings involving applicants from other origins. Latin American origins comprise the 33 member states of the United Nations Latin American and Caribbean Group, as defined in Section 5.3.4. Panel A.36a plots local linear regressions of the residualized lower-court reversal indicator on leave-out judge leniency. The reversal indicator equals one when a lower-court denial is followed by an appeal grant and zero otherwise. The solid black line represents Latin American origins, while the solid gray line represents other origins. The dashed black lines and gray shaded area report 95% confidence intervals based on standard errors clustered at the judge level. The gray bars show the distribution of judge leniency in the estimation sample. Panel A.36b reports the corresponding MTE functions over the common mean-centered grid of predicted lower-court grant probabilities, estimated as described in Section 5.3.2. The solid black and gray lines again represent Latin American and other origins, respectively. The dashed black lines and gray shaded area report pointwise 90% bootstrap percentile confidence intervals based on 500 replications resampled at the judge level.

Figure A.37: Reversal Rates and MTEs by Conflict-Affected Origin

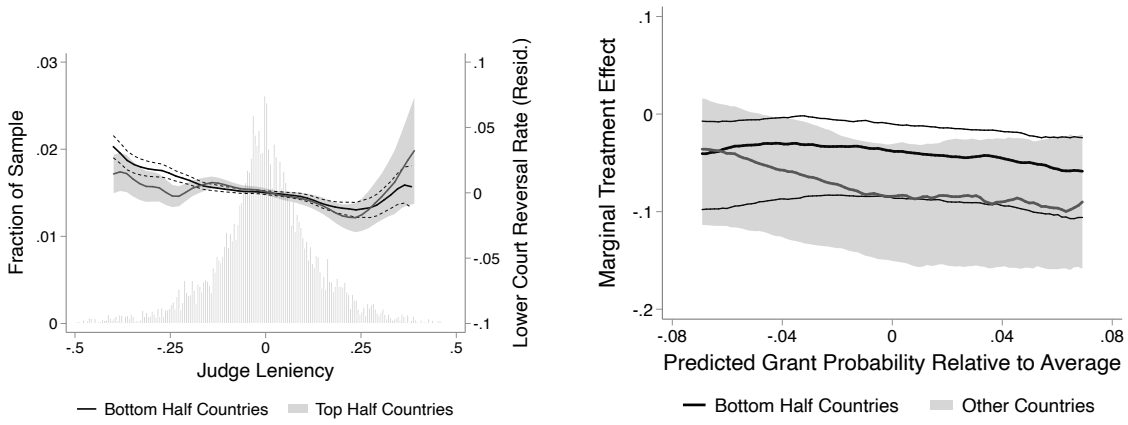


(a) Reduced Form

(b) Marginal Treatment Effect

Notes: This figure compares proceedings involving applicants from recently conflict-affected origins with proceedings involving applicants from other origins. An origin is classified as recently conflict-affected when the UCDP/PRIO Armed Conflict Dataset records at least one state-based armed conflict during the lower-court decision year or either of the two preceding years, as defined in Section 5.3.4. Panel A.37a plots local linear regressions of the residualized lower-court reversal indicator on leave-out judge leniency. The reversal indicator equals one when a lower-court denial is followed by an appeal grant and zero otherwise. The solid black line represents recently conflict-affected origins, labeled "Countries with Conflicts" in the graphic, while the solid gray line represents other origins. The dashed black lines and gray shaded area report 95% confidence intervals based on standard errors clustered at the judge level. The gray bars show the distribution of judge leniency in the estimation sample. Panel A.37b reports the corresponding MTE functions over the common mean-centered grid of predicted lower-court grant probabilities, estimated as described in Section 5.3.2. The solid black and gray lines again represent recently conflict-affected and other origins, respectively. The dashed black lines and gray shaded area report pointwise 90% bootstrap percentile confidence intervals based on 500 replications resampled at the judge level.

Figure A.38: Reversal Rates and MTEs by Origin-Country GDP per Capita



(a) Reduced Form

(b) Marginal Treatment Effect

Notes: This figure compares proceedings involving applicants from lower-income origins with proceedings involving applicants from other origins. For each lower-court decision year, an origin is classified as lower income when its GDP per capita at purchasing-power parity in constant 2021 international dollars lies in the bottom half of the available country distribution. The construction and data sources are described in Section 5.3.4. Panel A.38a plots local linear regressions of the residualized lower-court reversal indicator on leave-out judge leniency. The reversal indicator equals one when a lower-court denial is followed by an appeal grant and zero otherwise. The solid black line represents bottom-half origins. The solid gray line represents top-half origins, labeled "Top Half Countries" in Panel A.38a and "Other Countries" in Panel A.38b. The dashed black lines and gray shaded area report 95% confidence intervals based on standard errors clustered at the judge level. The gray bars show the distribution of judge leniency in the estimation sample. Panel A.38b reports the corresponding MTE functions over the common mean-centered grid of predicted lower-court grant probabilities, estimated as described in Section 5.3.2. The dashed black lines and gray shaded area in that panel report pointwise 90% bootstrap percentile confidence intervals based on 500 replications resampled at the judge level.

Table A.12: Average Marginal Treatment Effects by Origin Group, Common Trimmed Support

	(1) All	(2) Attentive Judges	(3) Inattentive Judges
Panel A: Latin American versus other origins			
$\overline{\text{MTE}}_{\text{latin}}$	-0.0060	-0.0255	-0.3409
	[-0.1423, 0.1000]	[-0.5855, 0.2507]	[-1.1875, 1.1670]
$\overline{\text{MTE}}_{\text{other}}$	-0.1556	-0.1973	-0.3648
	[-0.2808, -0.0309]	[-0.6143, 0.1054]	[-0.9677, -0.1913]
$\overline{\text{MTE}}_{\text{latin}} - \overline{\text{MTE}}_{\text{other}}$	0.1496	0.1718	0.0239
	[-0.0280, 0.3315]	[-0.4206, 0.7162]	[-0.8851, 1.6805]
sup-t test	1.5457	1.4607	0.5994
global bootstrap p-value	0.5250	0.4431	0.7944
Panel B: Recently conflict-affected versus other origins			
$\overline{\text{MTE}}_{\text{conflict}}$	-0.1007	-0.0784	-0.3356
	[-0.1493, -0.0178]	[-0.2179, 0.0951]	[-0.5207, 0.0247]
$\overline{\text{MTE}}_{\text{other}}$	-0.0262	-0.0901	-0.1457
	[-0.0509, 0.0178]	[-0.1610, 0.0534]	[-0.2618, 0.0206]
$\overline{\text{MTE}}_{\text{conflict}} - \overline{\text{MTE}}_{\text{other}}$	-0.0745	0.0117	-0.1899
	[-0.1428, 0.0071]	[-0.1980, 0.1863]	[-0.4272, 0.1920]
sup-t test	1.7789	0.4291	1.2277
global bootstrap p-value	0.2595	0.9940	0.3892
Panel C: Bottom-half versus top-half GDP-per-capita origins			
$\overline{\text{MTE}}_{\text{bottom}}$	-0.0415	-0.0831	-0.1810
	[-0.0663, 0.0049]	[-0.1657, 0.0426]	[-0.2755, -0.0223]
$\overline{\text{MTE}}_{\text{other}}$	-0.0730	-0.0192	-0.3614
	[-0.1279, -0.0089]	[-0.2181, 0.2724]	[-0.5638, -0.0166]
$\overline{\text{MTE}}_{\text{bottom}} - \overline{\text{MTE}}_{\text{other}}$	0.0315	-0.0639	0.1803
	[-0.0320, 0.1068]	[-0.3510, 0.1902]	[-0.1673, 0.4722]
sup-t test	1.1324	0.6132	1.1797
global bootstrap p-value	0.6228	0.8383	0.4351

Notes: This table reports uniformly weighted average marginal treatment effects (MTEs) over the full common trimmed support for the origin-group comparisons defined in Section 5.3.4. The estimation sample consists of lower-court denials, and the outcome equals one when a denial is followed by an appeal grant and zero otherwise. Column (1) pools all judges; Columns (2) and (3) report attentive and inattentive judges, defined as the top and bottom terciles of grant-rate response magnitude within ruling year. Panels A, B, and C compare Latin American origins, recently conflict-affected origins, and bottom-half GDP-per-capita origins with their respective comparison groups. Each average MTE integrates the estimated MTE function over the full common trimmed support of the mean-centered predicted grant-probability grid and divides that integral by the width of the support, following Section 5.3.2. Each difference row reports the first listed group minus the comparison group. Brackets report 90% reverse-percentile confidence intervals based on 500 judge-level bootstrap replications. Within each panel and column, the sup-t statistic and global bootstrap p-value test equality of the two MTE functions over the full common trimmed support.

Table A.13: Average Marginal Treatment Effects by Origin Group, Strict Region

	(1) All	(2) Attentive Judges	(3) Inattentive Judges
Panel A: Latin American versus other origins			
$\overline{\text{MTE}}_{\text{latin}}$	-0.0084	0.0274	-0.1048
	[-0.0645, 0.0527]	[-0.1764, 0.1827]	[-0.3635, 1.2363]
$\overline{\text{MTE}}_{\text{other}}$	-0.1780	-0.1457	-0.3845
	[-0.3580, -0.0202]	[-0.9746, 0.1709]	[-1.1531, -0.2271]
$\overline{\text{MTE}}_{\text{latin}} - \overline{\text{MTE}}_{\text{other}}$	0.1696	0.1731	0.2797
	[-0.0032, 0.3726]	[-0.2074, 0.9917]	[-0.0388, 1.9927]
sup-t test	1.5457	0.8834	0.5994
global bootstrap p-value	0.3114	0.7804	0.7784
Panel B: Recently conflict-affected versus other origins			
$\overline{\text{MTE}}_{\text{conflict}}$	-0.0981	-0.1127	-0.3775
	[-0.1630, -0.0069]	[-0.2798, 0.0721]	[-0.6015, 0.0182]
$\overline{\text{MTE}}_{\text{other}}$	-0.0141	-0.1000	-0.1916
	[-0.0398, 0.0327]	[-0.1865, 0.0570]	[-0.3371, -0.0103]
$\overline{\text{MTE}}_{\text{conflict}} - \overline{\text{MTE}}_{\text{other}}$	-0.0840	-0.0127	-0.1859
	[-0.1675, 0.0048]	[-0.2498, 0.1907]	[-0.4717, 0.2681]
sup-t test	1.7789	0.3903	1.0566
global bootstrap p-value	0.1697	0.9741	0.4291
Panel C: Bottom-half versus top-half GDP-per-capita origins			
$\overline{\text{MTE}}_{\text{bottom}}$	-0.0347	-0.0924	-0.2211
	[-0.0620, 0.0186]	[-0.1851, 0.0546]	[-0.3455, -0.0421]
$\overline{\text{MTE}}_{\text{other}}$	-0.0570	-0.0108	-0.4067
	[-0.1087, 0.0108]	[-0.2421, 0.2766]	[-0.6621, -0.0575]
$\overline{\text{MTE}}_{\text{bottom}} - \overline{\text{MTE}}_{\text{other}}$	0.0224	-0.0817	0.1856
	[-0.0456, 0.1011]	[-0.3750, 0.2252]	[-0.2125, 0.5042]
sup-t test	1.1324	0.6132	0.9350
global bootstrap p-value	0.4451	0.6407	0.5150

Notes: This table reports uniformly weighted average marginal treatment effects (MTEs) over the strict region for the origin-group comparisons defined in Section 5.3.4. The estimation sample consists of lower-court denials, and the outcome equals one when a denial is followed by an appeal grant and zero otherwise. Column (1) pools all judges; Columns (2) and (3) report attentive and inattentive judges, defined as the top and bottom terciles of grant-rate response magnitude within ruling year. Panels A, B, and C compare Latin American origins, recently conflict-affected origins, and bottom-half GDP-per-capita origins with their respective comparison groups. The strict region is the lower half of the common trimmed support and corresponds to the strict-judge region emphasized in Section 5.1. Each difference row reports the first listed group minus the comparison group. Brackets report 90% reverse-percentile confidence intervals based on 500 judge-level bootstrap replications. Within each panel and column, the sup-t statistic and global bootstrap p-value test equality of the two MTE functions over the strict region.

B Prediction Models and Evaluation

This section reports implementation details for the lower-court and BIA random-forest models and presents additional evaluation results for the BIA classifier. The lower-court model measures early predictability from predetermined observables, while the BIA model classifies realized appellate feedback by its diagnostic content. We first describe the common preprocessing, tuning, and sequential prediction procedures, then present additional evidence on the performance and predictive features of the BIA classifier. The lower-court exercise extends the prediction framework of Dunn et al. (2017).

B.1 Model Training and Implementation

This subsection describes the data preprocessing, hyperparameter tuning, and sequential prediction procedures used for the lower-court and BIA models.

B.1.1 Data Preprocessing Pipeline

We apply a common data-preprocessing pipeline to the lower-court proceeding sample and the binary BIA prediction sample. The pipeline has three steps. First, we impute missing values using variable-specific methods, including explicit handling of numerical features without a predefined imputation rule. Second, we convert numerical and categorical variables to consistent data types. Third, we one-hot encode categorical variables using fixed sets of possible values so that the training and prediction data share the same feature space. For aggregate-sample tuning and evaluation, we apply these transformations before creating a reproducible random training-test partition. This partition is separate from the sequential prediction design used in the empirical analysis, which trains each annual model on earlier filing cohorts and predicts the current filing cohort. The same pipeline provides consistent imputation, data types, and feature encoding across the two prediction exercises.

B.1.2 Hyperparameter Tuning

For the lower-court prediction, we conduct a grid search using three-fold stratified cross-validation. The parameter grid includes maximum tree depth values of

30 and unrestricted depth; 50, 100, 200, or 400 trees; minimum split sizes of 2 or 5; and minimum leaf sizes of 1 or 2. We select hyperparameters by maximizing ROC AUC. The selected specification uses 400 trees and a minimum split size of 5. Tuning is performed once on the aggregate lower-court prediction sample, and the selected hyperparameters are then applied to the sequential models and the five early-predictability specifications.

For the BIA prediction, we use the same tuning procedure with maximum tree depth values of 20, 40, and unrestricted depth; 50, 100, 200, or 400 trees; minimum split sizes of 2 or 5; and minimum leaf sizes of 1, 2, or 5. The selected specification uses 400 trees, a minimum split size of 2, and a minimum leaf size of 1. Tuning is again performed once on the aggregate binary BIA prediction sample, after which the selected hyperparameters are applied to the annual sequential models. This aggregate-sample tuning exercise is distinct from the out-of-sample sequential evaluation reported in Section 3.2 and Figure B.3. To benchmark aggregate-sample performance, we compare random forests with Extra Trees, Gradient Boosting, Histogram-based Gradient Boosting, and Logistic Regression. Figure B.1 shows that Random Forest and Extra Trees lead the comparison and perform similarly: Extra Trees achieves a ROC AUC of 0.856, compared with 0.848 for Random Forest, and both exceed the remaining benchmarks. Given its comparable performance and its suitability for the categorical, nonlinear, and interactive relationships described in Section 3, we use the random-forest model for both prediction exercises.

B.1.3 Sequential Prediction Design

For both prediction exercises, filing year defines the annual prediction cohorts. For a cohort filed in year t, the model is trained using observations filed before year t and then applied to the cohort filed in year t. The observations are lower-court proceedings in the early-predictability exercise and appeal records in the BIA exercise. This ordering keeps the current and later filing cohorts out of the corresponding model fit.

We allow the predictive relevance of earlier observations to decline with their temporal distance from the prediction cohort. Figure B.2 documents substantial variation in appeal grant rates and appeal volume across filing cohorts, motivating an exponential weighting scheme that places less weight on older training obser-

vations. Formally, let $\tau \geq 1$ denote the number of years between the prediction cohort and a training observation. We assign that observation weight

$$\alpha = \delta^{\tau-1},$$

where $\delta \in [0, 1]$ controls the rate of decay. A value of $\delta = 1$ assigns equal weight to all earlier filing cohorts, while lower values place progressively less weight on older observations.

We select δ separately for the two prediction exercises using ROC AUC as the performance criterion. We first search over a coarse grid from 0.1 to 1.0 in increments of 0.1 and then search over a 0.01 grid around the best-performing value from the first stage. The selected decay parameter is 0.90 for the lower-court model and 0.89 for the BIA model.

B.2 BIA Prediction Evaluation

Figure B.3 compares the accuracy of annual classification across the aggregate, unweighted sequential, and optimally weighted sequential BIA models. The aggregate model generally achieves higher accuracy because it is fitted using observations from across the full sample period, whereas each sequential model is restricted to earlier filing cohorts. The weighted and unweighted sequential estimates track each other closely, indicating that recency weighting makes little difference to annual classification accuracy. The optimally weighted sequential model remains the preferred classifier used in the main analysis because it preserves the filing-cohort ordering while allowing the predictive relevance of earlier observations to decline.

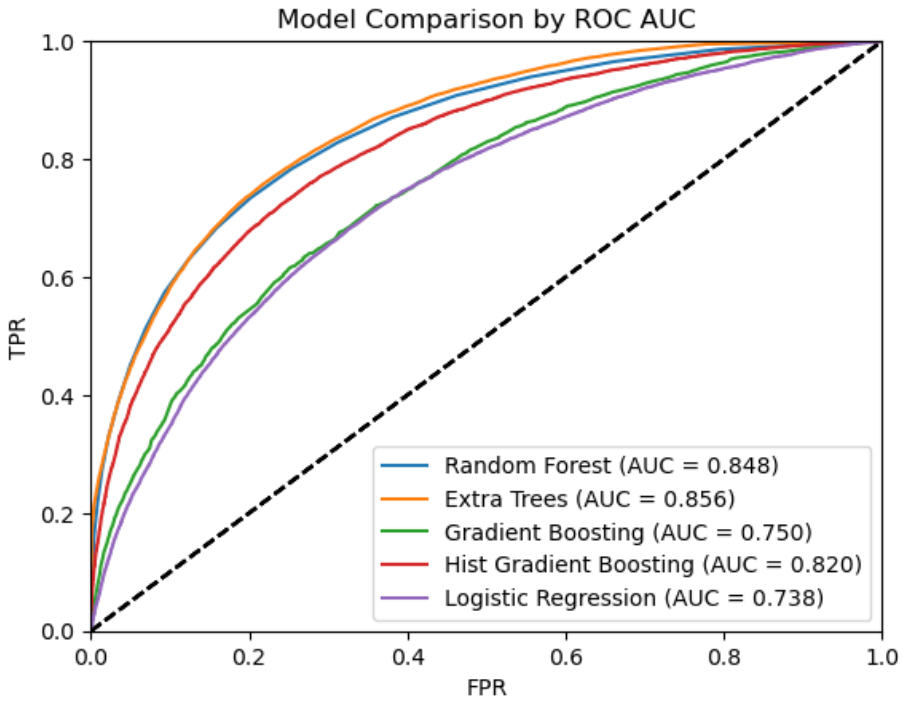
Predictive performance also varies across filing cohorts. The sharpest decline occurs in 2009 and is concentrated among appeals filed by Haitian applicants. As Figure B.4 shows, classification accuracy for Haitian appeals falls sharply in 2009, while accuracy for applicants from other countries changes much less. This timing coincides with the January 2010 Haitian earthquake, illustrating how an unforeseen country-specific shock can reduce the predictive value of earlier filing cohorts.

Figure B.5 compares feature importance across the aggregate, unweighted sequential, and optimally weighted sequential BIA models. Across the three models, the most predictive features include the time between the lower-court decision

and appeal filing, applicant nationality, and lower-court decision timing. Applicant nationality has the greatest mean importance in both sequential models. The aggregate model places relatively greater importance on calendar-time features, consistent with its use of observations from across the full sample period, while the sequential models use only earlier filing cohorts. Feature importance describes each variable's contribution to prediction within the fitted models and does not have a causal interpretation.

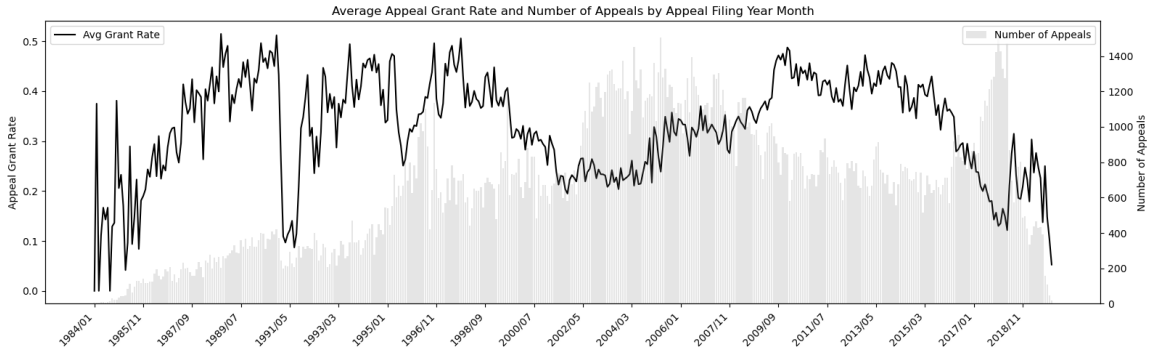
Table B.1 groups the predictors used in the sequential BIA model into six substantive categories. Time-horizon features have the largest grouped importance, followed by judge-related and respondent features. These values summarize the predictive importance assigned to the listed variables within the fitted model; they do not isolate the causal influence of immigration judges or other characteristics on BIA dispositions.

Figure B.1: Aggregate-Sample BIA Model Comparison



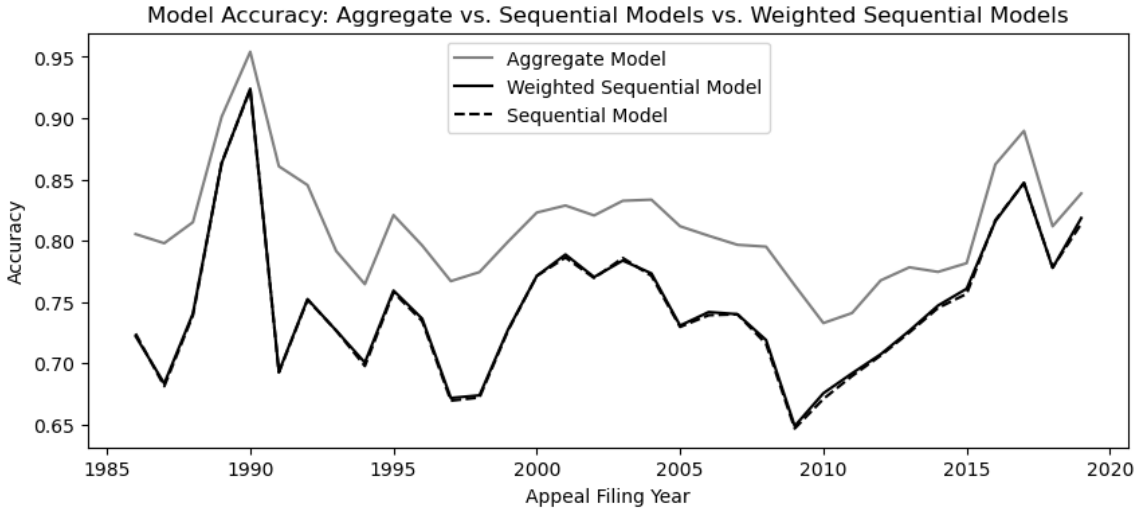
Notes: The figure plots receiver operating characteristic (ROC) curves for five classification algorithms trained and evaluated using the aggregate binary BIA prediction sample. The legend reports the area under each ROC curve. This aggregate-sample comparison is distinct from the sequential BIA prediction design used in the empirical analysis.

Figure B.2: Monthly Appeal Grant Rate and Volume, 1984-2019



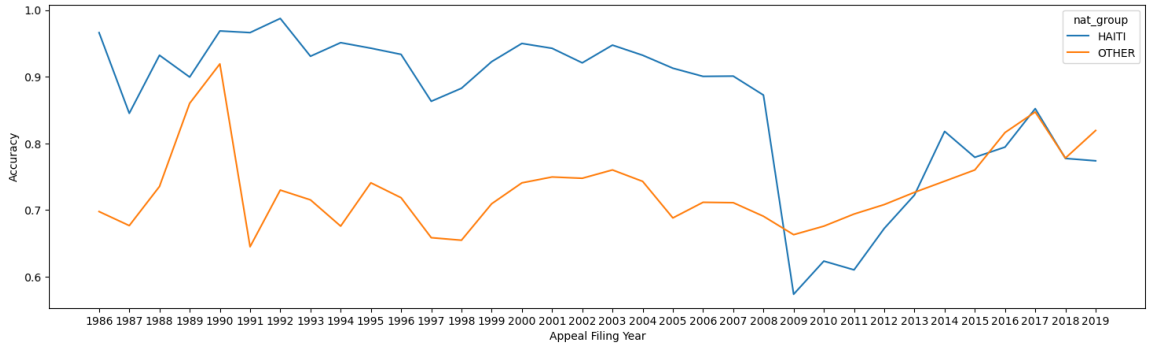
Notes: The figure reports the monthly appeal grant rate and number of appeal records in the binary BIA prediction sample, grouped by appeal filing month. The solid line reports the share of records classified as appeal grants, and the gray bars report the number of appeal records. Appeal grant is shorthand for the positive BIA disposition category defined in Appendix Table A.3.

Figure B.3: Annual Accuracy of Aggregate and Sequential BIA Models, 1986-2019



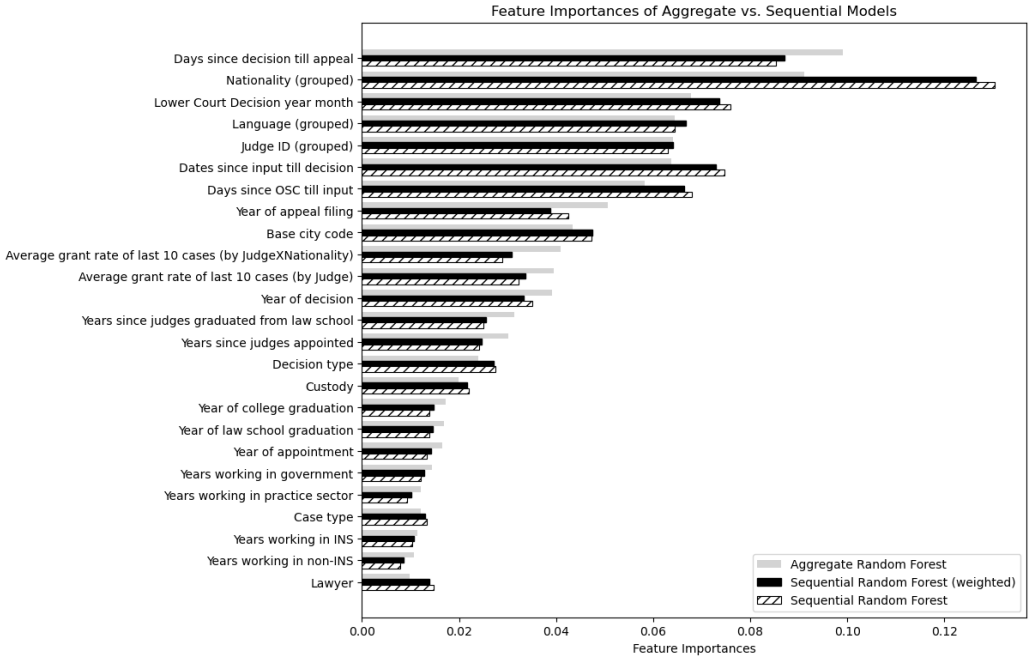
Notes: The figure reports classification accuracy within each appeal filing-year cohort in the binary BIA prediction sample. Accuracy is the share of appeal records correctly classified as appeal grants or appeal denials. The gray line reports the aggregate model, the dashed black line reports the unweighted sequential model trained on earlier filing cohorts, and the solid black line reports the optimally weighted sequential model using the decay parameter defined above.

Figure B.4: Annual BIA Prediction Accuracy for Haitian and Other Applicants, 1986-2019



Notes: The figure reports the accuracy of the optimally weighted sequential BIA model within each appeal filing-year cohort, separately for records involving Haitian applicants and records involving applicants from all other countries. The sample is the binary BIA prediction sample, and accuracy is the share of appeal records correctly classified as appeal grants or appeal denials.

Figure B.5: Feature Importance across BIA Prediction Models



Notes: The figure reports relative random-forest feature importance for the aggregate BIA model and mean feature importance across the annual unweighted and optimally weighted sequential models. Gray bars represent the aggregate model, solid black bars represent the optimally weighted sequential model, and hatched bars represent the unweighted sequential model. Feature importance measures predictive contributions within the fitted models and does not represent causal effects on BIA dispositions.

Table B.1: Grouped Feature Importance in the Sequential BIA Model

Feature Group	**Feature Importance**	**Features Included**
Time-horizon features	0.375	lower-court decision year-month, lower-court decision year, appeal filing year-month, appeal filing year, 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	0.263	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
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, pro bono
Location features	0.045	base city, indicator of hearing locations matching with base cities

Notes: The table reports grouped mean feature importance from the optimally weighted annual sequential random-forest BIA models. The third column lists the mutually exclusive features assigned to each group. The grouped importance values sum to one, subject to rounding, and describe predictive contributions within the fitted models.

C Proof of Proposition 5.1 and Lemma 5.1

Proof of Proposition 5.1 For $c > c^*$, under the assumption that $f(c) > 0$ and $c'(l) < 0$, the unconditional and denial-conditioned reversal rates are

$$R(c) = \int_{c^*}^{c} f(q)\,\mathrm{d}q, \qquad R^D(c) = \frac{F(c) - F(c^*)}{F(c)}.$$

The Fundamental Theorem of Calculus and the chain rule give

$$\frac{\mathrm{d}R(c(l))}{\mathrm{d}l} = f(c)c'(l) < 0, \qquad \frac{\mathrm{d}R^D(c(l))}{\mathrm{d}l} = \frac{F(c^*)f(c)}{F(c)^2}c'(l) < 0.$$

The selection margin is

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

Therefore,

$$\frac{\mathrm{d}E(c(l))}{\mathrm{d}l} = \frac{F(a)f(c)}{F(c)^2}c'(l) \leq 0,$$

with strict inequality whenever $F(a) > 0$.

The success margin is

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

Its derivative with respect to leniency is

$$\frac{\mathrm{d}I(c(l))}{\mathrm{d}l} = \frac{f(c)\left[F(c^*) - F(a)\right]}{\left[F(c) - F(a)\right]^2}c'(l) < 0.$$

■

Proof of Lemma 5.1 The sensitivity functions are:

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

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

$$\begin{aligned}
\frac{\mathrm{d}\psi_R(c)}{\mathrm{d}c} &= f'(c) > 0, \\
\frac{\mathrm{d}\psi_{R^D}(c)}{\mathrm{d}c} &= F(c^*)\frac{f'(c)F(c) - 2f(c)^2}{F(c)^3} > 0, \\
\frac{\mathrm{d}\psi_E(c)}{\mathrm{d}c} &= F(a)\frac{f'(c)F(c) - 2f(c)^2}{F(c)^3} \geq 0.
\end{aligned}$$

Thus, $\psi_R(c)$ and $\psi_{R^D}(c)$ are strictly increasing in c over this interval. $\psi_E(c)$ is non-decreasing and is strictly increasing whenever $F(a) > 0$.

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

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

Therefore, $\psi_I(c)$ is also strictly increasing in c over this interval. ■

Assumptions of Lemma 5.1 The regularity conditions in Lemma 5.1 are imposed over $[c^*, c_{j,0}]$:

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

These conditions govern how the sensitivity of the reversal outcomes changes with threshold strictness.

The first condition requires the case-quality density to increase over the relevant range and therefore makes the unconditional reversal sensitivity rise with c. The second requires the density to rise sufficiently quickly relative to the cumulative mass $F(c)$, making the denial-conditioned reversal sensitivity increase with c. It also makes the selection-margin sensitivity increase when $F(a) > 0$; when $a = 0$, all denied cases satisfy the appeal-filing cutoff and the selection margin is constant. The third condition makes the BIA-success sensitivity increase with c. When $F(a) > 0$, it is the stronger condition because $F(c) - F(a) < F(c)$. Economically, these restrictions place relatively more case density near stricter decision thresholds. Because a marginal case at $c > c^*$ also lies above the appeal-filing threshold, an increase in the lower-court threshold moves cases that would appeal and satisfy the BIA standard into the denial pool. Under the stated conditions, reversal outcomes are therefore more sensitive to threshold changes at the initial

strict threshold than near the BIA-aligned threshold.

The model also establishes the ordering used in the MTE analysis. Because

$$\Pr(D_i = 1 \mid l) = 1 - F(c(l)),$$

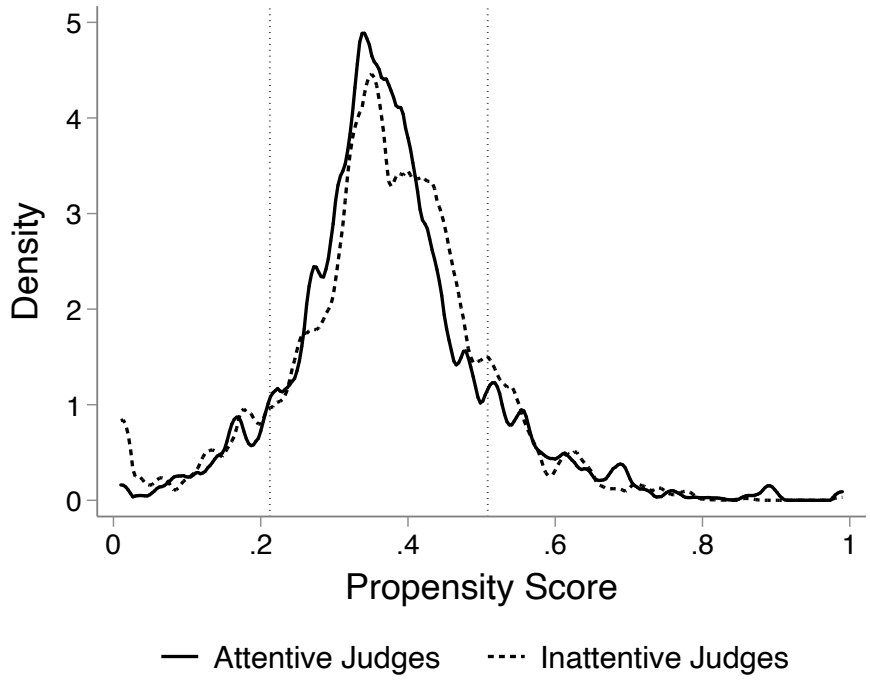
we have

$$\frac{\mathrm{d}}{\mathrm{d}l} \Pr(D_i = 1 \mid l) = -f(c(l))c'(l) > 0.$$

The model-implied grant probability therefore rises with judge leniency and falls with threshold strictness. The empirical predicted grant probability p is the first-stage counterpart to this ordering. It indexes the component of granting predicted by judge leniency and is not the latent case quality q.

The regularity conditions in Lemma 5.1 concern the latent case-quality density. Figure C.1 serves a separate empirical purpose by documenting support for the MTE comparisons. It plots the full-sample distribution of predicted lower-court grant probabilities separately for attentive and inattentive judges. The two distributions overlap over the common trimmed region bounded by the 10th and 90th percentiles of the full-sample distribution. This region defines the fixed evaluation grid used for both groups and across bootstrap replications.

Figure C.1: Distribution of Predicted Lower-Court Grant Probabilities



Notes: The figure plots kernel-density estimates of the full-sample predicted lower-court grant probability described in Section 5.3.2, separately for proceedings ruled by attentive and inattentive judges. Attentive and inattentive judges correspond to the top and bottom terciles of grant-rate response magnitude within ruling year, as described in Section 5.2.1. The solid line represents attentive judges, and the dashed line represents inattentive judges. The vertical dotted lines mark the 10th and 90th percentiles of the full-sample distribution. Predicted probabilities are bounded between 0.01 and 0.99 before these percentiles are calculated. The percentile bounds define the common fixed grid of 200 evaluation points used for both groups and across bootstrap replications.

D Review Margins and the Reversal MTE

This section establishes the conditional-mean identity underlying Equation (5.8). To simplify notation, we suppress the case, judge, and time subscripts. Within the lower-court denial sample, let $A \in \{0,1\}$ indicate that a denial is appealed, and let $S \in \{0,1\}$ indicate an appeal grant among appealed denials. The denial-conditioned reversal outcome is $Y = A \cdot S$. Therefore,

$$E[Y \mid p] = E[A \mid p]E[S \mid A = 1, p].$$

This level identity states that the denial-conditioned reversal probability equals the appeal-filing probability multiplied by the BIA-success probability conditional on appeal.

The same identity holds within each court-time cell C:

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

Both expressions are exact identities for unresidualized conditional means.[35]

Differentiating Equation (D.1) with respect to p gives

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

This product rule describes the derivative of the unresidualized conditional mean within a fixed court-time cell.

For Y and A, define residualized outcomes within the lower-court denial sample as

$$\tilde{Y} = Y - E[Y \mid C] \qquad \tilde{A} = A - E[A \mid C].$$

[35] By the law of iterated expectations,

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

Within the appealed-denial sample, define

$$\tilde{S} = S - E[S \mid A = 1, C].$$

Because the corresponding cell mean is constant with respect to p within a fixed C, residualization leaves each within-cell derivative unchanged:

$$\begin{aligned}
\frac{\partial}{\partial p} E[\tilde{Y} \mid p, C] &= \frac{\partial}{\partial p} E[Y \mid p, C], \\
\frac{\partial}{\partial p} E[\tilde{A} \mid p, C] &= \frac{\partial}{\partial p} E[A \mid p, C], \\
\frac{\partial}{\partial p} E[\tilde{S} \mid A = 1, p, C] &= \frac{\partial}{\partial p} E[S \mid A = 1, p, C].
\end{aligned}$$

Define the within-cell residualized MTEs as

$$\begin{aligned}
\text{MTE}_{Y,C}(p) &\equiv \frac{\partial}{\partial p} E[\tilde{Y} \mid p, C], \\
\text{MTE}_{A,C}(p) &\equiv \frac{\partial}{\partial p} E[\tilde{A} \mid p, C], \\
\text{MTE}_{S,C}(p) &\equiv \frac{\partial}{\partial p} E[\tilde{S} \mid A = 1, p, C].
\end{aligned}$$

The within-cell product rule can then be written as

$$\begin{aligned}
\text{MTE}_{Y,C}(p) = &\underbrace{\text{MTE}_{A,C}(p) \cdot E[S \mid A = 1, p, C]}_{\text{selection-margin term}} \\
&+ \underbrace{E[A \mid p, C] \cdot \text{MTE}_{S,C}(p)}_{\text{success-margin term}}.
\end{aligned}$$

The empirical analysis instead estimates pooled residualized MTEs for $\tilde{Y}$, $\tilde{A}$, and $\tilde{S}$ separately on their relevant samples. The appeal-filing and BIA-success MTEs describe two distinct margins of the review process. Because the displayed curves are not the weighted within-cell terms above, they do not mechanically add to the displayed reversal MTE.