Weighted Noise: Discretion in Regulation

Sumit Agarwal, Bernardo Morais, Amit Seru and Kelly Shue*

July 31, 2025

Abstract

Human discretion is a defining feature of how legal and regulatory institutions make complex decisions. While discretion lets professionals incorporate subtle soft information that rules or algorithms may miss, it also introduces noise—disagreement among decision-makers, such that the same case could yield different outcomes depending on who handles it. We study this trade-off in U.S. bank supervision, where examiners issue CAMELS ratings that shape bank capitalization and lending. Using confidential supervisory data, we model final ratings as weighted sums of component issues. We find that disagreement arises through three channels: (1) examiners treat all issues, even relatively objective ones, as subjective; (2) they disproportionately weight more subjective issues, with management quality alone accounting for nearly 50% of the total weight; and (3) they disagree on the weights applied to each issue, even when agreeing on the issues themselves. We measure discretion as the residual in ratings unexplained by bank fundamentals—this residual captures both soft information and examiner-specific judgment. Because examiner assignment is quasi-random, systematic variation in this residual isolates noise rather than signal. This noise has real consequences: tougher examiners cause banks to raise capital and cut lending for years, and the expectation of unpredictable supervision can prompt banks to preemptively reduce lending. Yet discretion is not purely negative; the discretionary component of ratings predicts future bank deterioration more accurately than a simple model based solely on hard data. Nonetheless, excessive discretion adds noise without extra signal. Moderately constraining how subjective components are weighted can improve predictive accuracy and reduce costly disagreement.

Keywords: noise, discretion, algorithm aversion, decision-making, regulation

*We thank Malcolm Baker, Nick Barberis, Jonathan Berk, John Cochrane, James Choi, Xavier Gabaix, Claudia Robles-Garcia, Matt Gentzkow, Justin Grimmer, Stephen Haber, Chad Jones, Pete Klenow, Eddie Lazear, Ross Levine (discussant), Lasse Pedersen, Paola Sapienza, David Scharfstein, Andrei Shleifer, Jeremy Stein, Adi Sunderam, Richard Thaler, Ali Yurukoglu, Luigi Zingales, Jeff Zwiebel, and participants in the Econometric Society Meetings (Behavioral Finance), Financial Literacy Colloquium at Stanford GSB, Banks and Beyond Conference at Hoover Institution as well as seminars at the American Enterprise Institute, Harvard Business School, Hoover Institution (Economics and Politics Seminar), Stanford GSB, and Yale SOM for helpful comments. Sumit Agarwal: Finance Department, NUS. Bernardo Morais, Board of Governors of the Federal Reserve System. Amit Seru, Stanford University Graduate School of Business, Hoover Institution and NBER. Kelly Shue: Yale School of Management and NBER. We thank Winston Xu and Yiyang Han for excellent research assistantship. The views expressed in this paper are those of the authors and do not reflect the views of the Board of Governors or the Federal Reserve System. An earlier version of this paper was titled, "Noisy Experts? Discretion in Regulation." First Version: November 2019.

I Introduction

If two felons who both should be sentenced to five years in prison receive sentences of three years and seven years, justice has not, on average, been done. In noisy systems, errors do not cancel out. They add up.

- Kahneman, Sibony, and Sunstein (2021)

Institutions across society rely on human discretion to make consequential decisions. Judges, medical specialists, scientific referees, patent examiners, HR committees, and bank supervisors are deliberately given latitude to apply judgment. The promise of discretion is clear: humans can extract and interpret subtle "soft information" that rules or algorithms may overlook (Petersen and Rajan, 1994; Petersen and Rajan, 2002; Stein, 2002; Liberti and Mian, 2009; Rajan, Seru, and Vig, 2015). But the peril is equally real: when different professionals interpret the same facts differently, the resulting noise—inconsistency that does not wash out but systematically accumulates—can distort outcomes away from optimal benchmarks (Kahneman, Sibony, and Sunstein, 2021).

This tension is particularly acute in prudential banking supervision. U.S. bank examiners issue "CAMELS" ratings, composite scores summarizing a bank's safety and soundness based on six components: Capital, Assets, Management, Earnings, Liquidity, and Sensitivity to market risk. These ratings influence bank capitalization, lending, insurance costs, access to e okaymergency funding, and approvals for licenses or mergers. Examiners retain wide discretion, especially when assessing management quality, which relies heavily on interviews and subjective impressions. How much disagreement does this discretion generate? Where does it come from? Does it enhance oversight or undermine it by introducing costly noise? This paper investigates these questions.

The potential for discretion to create costly noise is underscored by research in psychology showing that humans often overweight subjective cues, sometimes underperforming very simple statistical models (Meehl, 1954; Dawes, 1979; Dawes, Faust, and Meehl, 1989; Huang and Pearce, 2015).[1] Recent advances in machine learning imply algorithmic forecasts have become increasingly difficult to outperform (Kleinberg et al., 2017; Hoffman, Kahn, and Li, 2018; Mullainathan and Obermeyer, 2022; Angelova, Dobbie, and Yang, 2022; Fuster et al., 2022). Yet algorithm aversion persists: people favor human judgment, even when it predictably introduces error. In regulation, this tension creates a double-edged sword: discretion may surface subtle signals but also exposes firms to unpredictable shocks and uneven treatment.

Despite its central role, systematic empirical evidence on how discretion generates disagreement and affects real outcomes is limited, especially in high-stakes regulatory domains like bank supervision. Most prior work focuses on judges and medical settings (e.g., Ramji-Nogales, Schoenholtz, and Schrag, 2010; Yang, 2015). We study a setting where discretion directly affects availability of credit and financial stability: routine bank safety-and-soundness exams in the U.S.

[1] Human judgment can also lead to discrimination against disadvantaged minorities (e.g., Yang, 2015; Kleinberg et al., 2018b; Benson, Li, and Shue, 2021; Arnold, Dobbie, and Yang, 2018; Arnold, Dobbie, and Hull, 2022).

We define noise as the disagreement that would occur if multiple decision makers considered the same case. We present a new framework and evidence showing disagreement stems from how individuals weight component issues. Final decisions can be modeled as weighted sums of ratings across these components. We identify three key sources of disagreement. First, decision-makers treat all factors, even relatively objective ones such as liquidity and capital adequacy, as if they were highly subjective. Second, they place disproportionate weight on the most subjective components. In our setting, examiners assign 50% of the weight to subjective assessments of management quality. Third, decision-makers differ in how they weight components, leading to disagreement in overall ratings even when they agree on individual parts. We show that examiner discretion has a large and persistent causal effect on future bank capitalization and credit supply. This contributes to volatility and uncertainty in outcomes, prompting banks to respond conservatively in anticipation. While replacing discretion with a simple algorithm worsens predictive accuracy, moderate limits on discretion improve the signal quality of supervisory ratings.

We use micro-data on decisions made by bank examiners during on-site "safety and soundness" examinations of U.S. banks. In each exam, a lead examiner and team review bank documents, assess the loan portfolio, and meet with management. The process culminates in a composite CAMELS rating, summarizing the bank's overall condition. This composite rating is based on six component ratings—Capital, Assets, Management, Earnings, Liquidity, and Sensitivity to market risk—collectively forming the CAMELS acronym. Examiners exercise discretion in assessing each component and in aggregating them into the composite score. Ratings range from 1 to 5, with higher scores indicating greater concern. CAMELS ratings carry significant implications. They influence a bank's FDIC insurance premiums, access to the Federal Reserve's discount window, and approvals for licensing, branching, and mergers. Banks typically respond to higher ratings by raising capital and cutting lending.

We focus our analysis on banks subject to regular examiner rotation, which together hold 6.5 trillion in assets—approximately 32% of total assets in the sector. While the largest banks are excluded due to their exemption from rotation, our sample includes major regional banks such as Silicon Valley Bank (SVB) and First Republic Bank, whose 2023 failures sparked a regulatory crisis. Banks in our sample are rotated across examiners within their state or region, subject to examiner workload constraints (Agarwal et al., 2014). Many parts of our analysis rely on the identification assumption that, within a state and time period, the assignment of lead examiners is unrelated to a bank's underlying quality. We support this assumption empirically: an examiner's general leniency (measured excluding the current exam) is uncorrelated with both observable indicators of bank health and the future performance of the bank's loan portfolio originated prior to the examination.

We begin by measuring discretion in each examination as the residual portion of the rating that cannot be explained by observable bank characteristics. This residual captures both soft information about the bank and the examiner's personal idiosyncrasies in interpreting and synthesizing all available information—both soft and hard.[2] Because higher ratings indicate greater concern, a

[2]Soft information refers to real signals about bank quality that are difficult to quantify. We treat it as part

positive residual implies the examiner issued a tougher judgment than predicted by bank fundamentals, and a negative residual implies a more lenient one. A non-zero residual does not necessarily mean the decision was inappropriate; it may reflect the examiner's use of legitimate soft information. However, random assignment implies that, in expectation, the distribution of soft signals is identical across decision-makers. Therefore, any *predictable* variation in residuals across examiners—such as some examiners consistently giving higher residuals than others—captures noise introduced by the decision-maker.

Using our exam-level measure of discretion, we construct two examiner-level metrics: directional and absolute discretion. These measures are related but conceptually distinct and lend themselves to different empirical tests. Directional discretion is the average signed value of discretion across all exams conducted by an examiner. It captures whether an examiner tends to be tougher or more lenient than peers. Absolute discretion is the average of the absolute values of discretion. It reflects how much an examiner relies on case-specific soft information, gut feelings, or personal judgment, regardless of direction. Importantly, an examiner can have zero directional discretion (no consistent bias toward leniency or severity) while still exhibiting high absolute discretion.

We find economically large levels of absolute discretion among examiners, as well as wide variation in directional discretion. Simulations suggest that healthy banks expected to receive a rating of 2 (the modal rating) face a 4.2% chance of being rated an unsatisfactory 3 or higher solely due to examiner assignment. Affected banks would face increased regulatory scrutiny, not because of fundamentals, but because of examiner discretion. Conversely, 5.0% of banks that would otherwise receive a 2 are instead rated a 1, potentially escaping appropriate oversight. These results underscore that supervisory noise may not cancel out. We further show that most year-over-year changes in bank ratings are driven by changes in examiner assignment rather than shifts in bank fundamentals, suggesting that many rating revisions reflect noise rather than changes in bank health.

Next, we examine examiner disagreement, which we define as the cross-sectional variance in composite ratings that would arise if different examiners reviewed the same bank. In a separate data sample, when multiple examiners assess the same bank simultaneously (concurrent exam sample), we observe disagreement directly. For our main sample however, such overlap does not occur. To estimate disagreement more broadly, we leverage the quasi-random rotation of examiners across banks within a region. Because examiners are assigned to banks with similar expected values of hard and soft information within a region and time period, systematic differences in their rating behavior reveal disagreement. Specifically, if one examiner consistently gives higher ratings—or ratings with greater variance—than others observing equivalent information, this reflects disagreement. We quantify this as the portion of an exam's rating that is predicted by the examiner's leave-out-mean directional or absolute discretion, controlling for region-time fixed effects.

To understand the sources of examiner disagreement, we model final decisions as weighted sums of component ratings. A key strength of our data is that we observe both the composite rating

of discretion because its use depends on human interpretation. Without discretion, soft information cannot be empirically observed, since its defining feature is that it is open to interpretation.

and all six component ratings, allowing us to directly investigate how examiner weighting decisions contribute to disagreement. Importantly, our estimates of these weights do not rely on assumptions about examiner assignment randomness or the measurement of bank fundamentals. We show that overall disagreement arises from three main factors.

First, examiners exhibit disagreement during concurrent exams on all component issues, even relatively objective ones like capital adequacy and liquidity. Moreover, across the full sample, examiners' persistent tendencies reliably predict their component-level assessments. For instance, examiners who are generally tougher on capital adequacy tend to issue tougher assessments in a given case, even after controlling for the bank's actual condition. This suggests that examiner-specific tendencies shape assessments, even for components that can be tied to objective data.

Second, we find that examiners place the most weight—about 50 percent—on the most subjective CAMELS component: management quality. Unlike other components such as capital adequacy, which can be measured with hard data, management assessments are based on personal interviews and interactions with bank leadership. Management ratings exhibit the highest level of disagreement in concurrent exams and are the least predictable using hard information from Call Report data. The heavy weighting on management quality aligns with psychology research showing that people often overvalue face-to-face interactions and cues like facial expressions and language (Levine, McCornack, and Park, 1999).

This focus of examiners on management quality is especially relevant in light of post-crisis calls to expand examiner discretion following the failure of SVB. Prior to its collapse, SVB received an unsatisfactory composite CAMELS rating of 3, driven primarily by the high weight on management rating of 3, while all other components were rated a 2.[3] To the extent that early warnings were desirable, SVB illustrates the potential value of weighting subjective assessments.[4]

Third, we show that examiners differ in how they weight individual components, leading to disagreement in final ratings even when they agree on component scores. There is substantial heterogeneity in weights across all component issues, with the greatest variation in the management rating—the same component that exhibits the most disagreement overall. This interaction further amplifies divergence in final decisions. We also find that certain types of examiners systematically place more weight on specific components: tougher examiners and those with higher absolute discretion assign significantly greater weight to management quality. Overall, this heterogeneity in weighting means that even consensus on component assessments may not translate into agreement on composite ratings, particularly when subjective components receive disproportionate emphasis.

After decomposing the sources of disagreement, we ask whether the noise from examiner discretion matters. We use the examiner rotation system to estimate its causal impact on bank behavior. We instrument each exam's rating with the examiner's leave-out-mean residual rating—

[3]See https://www.federalreserve.gov/publications/files/svb-review-20230428.pdf and https://www.banking.senate.gov/imo/media/doc/Kupiec%20Testimony%205-17-23.pdf.

[4]At the same time, as Fed Vice Chair Bowman recently emphasized (June 6, 2025), CAMELS ratings should reflect material financial risks, not be overly driven by subjective judgment—underscoring this paper's core finding that discretion adds value, but excess discretion could introduce costly noise. See https://www.federalreserve.gov/newsevents/speech/bowman20250606a.htm.

based on the idea that examiners who tend to assign high ratings to other banks are more likely to do so again, independent of the current bank's fundamentals.

Using this variation, we find that an exogenous one-point increase in ratings leads to a 0.27 standard deviation rise in capitalization and a 1.08 standard deviation decline in loan growth within one year.[5] These effects persist for at least two to three years, indicating that examiner discretion has durable real effects. Quasi-random examiner assignment thus generates meaningful volatility in bank outcomes, as capitalization and lending shift based on which examiner is assigned.

Beyond affecting bank behavior *ex post*, we find suggestive evidence that the anticipation of supervisory unpredictability leads banks to act conservatively in advance—holding more capital and curbing credit supply. Banks in states where examiners exhibit high absolute discretion or large variation in directional discretion—making ratings harder to predict—tend to preemptively hold more capital and reduce lending, even after accounting for fundamentals and past state-level ratings. These actions likely reflect efforts to avoid unfavorable outcomes if assigned a tough examiner in the future. Our analysis offers empirical support for the theoretical predictions in Repullo (2024), in which greater noise in bank supervision can reduce bank risk-taking. These patterns also mirror how macroeconomic regulatory uncertainty dampens investment (Bloom, 2009; Gissler, Oldfather, and Ruffino, 2016).

In the final part of the paper, we show that discretion is not all noise. The discretionary component of supervisory ratings predicts near-term deterioration in bank conditions and future downgrades better than a simple mechanical model. This shows that soft information improves forecasts when used well.[6] These findings also contrast with the strong negative claims in Meehl (1954) and Dawes (1979), which argued that human experts consistently underperform even simple unit-weighted models.

However, we show that examiners who exercise high discretion do not outperform their peers in predictive power—implying that beyond a certain point, additional discretion mainly adds noise. A simple reweighting of component scores that limits emphasis on subjective elements can outperform the official composite rating.[7] This approach mirrors U.S. federal sentencing guidelines, which limit but do not eliminate judicial discretion (Yang, 2015). Our results suggest that applying similar constraints in bank supervision could reduce noise and improve predictive power, contributing to the broader discussion on the governance of financial institutions (Laeven and Levine, 2009; Huber, 2021). Taken together, our findings underscore a nuanced trade-off: discretion enables the use of

[5]These causal effects contrast with the predictive—but non-causal—power of ratings. An endogenously higher rating reflects expectations of future trouble, so higher ratings today predict higher ratings later. In contrast, our IV-based exogenous variation is uncorrelated with fundamentals, so an exogenously higher (tougher) rating induces conservative responses from the bank that lower future ratings.

[6]We purposely do not compare examiner ratings to complex machine learning models. Given the rapid pace of AI development, such a test would quickly become outdated. Even if human judgment currently outperforms a machine model, this may not hold in the near future. Nonetheless, human discretion remains central to regulatory practice, with no serious proposals to replace bank examiners or judges with machines. Therefore, we focus on whether modest constraints on discretion can improve predictive power while reducing noise.

[7]Simply offering decision-makers the option to rely on algorithms may not be effective. Hoffman, Kahn, and Li (2018) show that hiring managers override algorithmic recommendations even without having better information.

soft information that rigid rules miss, but unbounded discretion amplifies disagreement in ways that distort outcomes and fuel uncertainty.

Our paper complements the large body of research showing that decision-making reflects psychological biases, mistaken beliefs, skill limitations, inherited traits, personal experiences, and agency incentives (e.g., Tversky and Kahneman, 1974; Chan, Gentzkow, and Yu, 2022; Bohren et al., 2023; Agarwal et al., 2014). While prior studies often isolate one such factor, we emphasize the substantial heterogeneity across individuals in the combination of influences shaping judgment. This heterogeneity can lead to disagreement and introduce costly noise. Rather than focusing on a single psychological mechanism, we show that disagreement arises from variation in how decision-makers weight issues—driven by heavy emphasis on subjective components, differences in weighting strategies, and the tendency to treat even objective factors as subjective.

Our analysis contributes to the literature on algorithm aversion and human decision-making biases in three ways. First, our empirical setting does not exhibit a potential sample selection problem present in most other field studies of discretion and bias in decision-making, (e.g., Kleinberg et al., 2017; Hoffman, Kahn, and Li, 2018; Benson, Li, and Shue, 2019). In those studies, the econometrician only observes outcomes for affirmative decisions. For example, we only observe future criminal activity if the judge grants parole. In contrast, banks in our setting almost always continue to exist after the examination. Second, our unique data on both component and composite ratings lets us examine how discretion over weighting can generate disagreement. Third, we show how human discretion interacts with the regulatory framework and can influence both the *ex post* and *ex ante* credit supply of U.S. banks. While behavioral biases among households are often used to justify tighter regulatory oversight, our findings suggest that regulators themselves may exhibit personal biases that meaningfully affect how oversight is applied.

Our study also offers a new perspective on the limitations of random assignment of cases to decision-makers, which many view as a hallmark of good institutional design. At best, random assignment promotes fairness by ensuring that no individual or entity systematically faces tougher or more lenient treatment.[8] However, random assignment alone does not solve the deeper problem: human decision-makers can produce noisy judgments that deviate from case merits (Lipsky, 2010). At the extreme, a regime in which case outcomes are determined by fair but arbitrary coin flips could deliver substantial welfare losses.[9]

Finally, our paper speaks to the broad literature on the optimal design of regulation and its enforcement (e.g., Shleifer and Vishny, 1999). In banking, much of this work has focused on regulatory arbitrage (e.g., Agarwal et al., 2014; Calomiris and Gorton, 1991; Calomiris, 2006; Garicano, 2012). We highlight that the design of human regulatory oversight—how much discretion

[8] Assessment of the quality of institutions has largely focused on the extent to which case assignment succeeds at being truly random (Hall, 2010; Chilton and Levy, 2015). See also https://blogs.wsj.com/law/2013/11/04/the-problem-with-not-so-random-case-assignment.

[9] Our estimates of the causal effects of examiner discretion connect to a broader literature using judge fixed effects to identify the causal impact of judicial decisions (e.g., Dobbie and Song, 2015; Bris, Welch, and Zhu, 2006; Chang and Schoar, 2006; Sampath and Williams, 2019). The strength of these instruments underscores how influential, and potentially distortionary, human discretion can be.

is permitted, structured, and bounded—should be as central to financial stability debates as the content of the rules themselves.

II Institutional background

II.A Related literature on CAMELS ratings

This paper is related to Agarwal et al. (2014), who find that federal bank examiners assign tougher CAMELS ratings than state examiners. Like Agarwal et al. (2014), we also exploit quasi-random rotation in examiner assignment as part of our identification strategy, but our focus and findings differ in several ways. First, Agarwal et al. (2014) zoom in on one specific reason driving one group of examiners to give more lenient ratings than another: state agencies offer incentives for more lenient ratings compared to federal agencies. In this paper, we take the view that state vs. federal agency incentives is just one factor out of the large set of potential psychological, inherited, experience, and agency factors that can influence examiner (federal or state) decision-making. Moreover, we investigate all forms of discretion, including absolute discretion which can manifest through increased uncertainty rather than predictable differences in means across examiners. Notably, we also show that, *even within the set of state or federal examiners*, there exists large individual variation in examiner discretion that is quantitatively larger than the cross-agency differences shown in Agarwal et al. (2014). Second, we study the determinants of disagreement among examiners and document, for the first time, the large role of subjective assessment of banks' management quality as well as heterogeneity in weights attached to more objective issues such as capital adequacy. Finally, we show that human discretion (when used within limits) can lead to more informative predictions of future bank health.

Our research is also related to a broader literature examining predictors of bank health. One branch of this literature compares the predictive power of public and private information. Given the wide-ranging consequences of bank failure, researchers and central bankers are particularly interested in predicting bank failure, and CAMELS have been shown to be central in such forecasting exercises and bank governance (e.g., Hirtle and Lopez, 1999; Berger et al., 2000; Calomiris, 2006; Laeven and Levine, 2009).

II.B An overview of US banking supervision

Bank supervision in the United States relies on two main pillars: off- and on-site monitoring. Off-site monitoring requires all depository institutions to file quarterly "Reports of Condition and Income," or Call Reports. Regulators use Call Reports to monitor a bank's financial condition between on-site examinations. On-site "safety and soundness" examinations are used to verify the content of Call Reports and to gather additional in-depth information regarding the safety and soundness of the supervised entity as well as its compliance with regulations. Since the Federal Deposit Insurance Corporation Improvement Act of 1991, bank examiners are required to conduct

on-site examinations every 12 months, unless bank assets fall below a minimum threshold, in which case the exams are conducted every 18 months.[10]

In an on-site examination, examiners read additional documents from the bank, review and evaluate its loan portfolio, and meet with the bank's management. Examiners comment on areas that must be improved; and depending on the bank's condition, they also discuss with management the need for informal or formal supervisory actions. Informal actions are established through a commitment from the bank to solve the deficiencies identified in the form of a memorandum of understanding or a bank board resolution. Formal actions are more severe. They include cease-and-desist orders, suspensions or removals of banks' senior management, and terminations of insurance.

These examinations culminate in the assignment of a CAMELS rating by a lead examiner, which summarizes the conditions of the bank (broken down into six components: capital adequacy, asset quality, management, earnings, liquidity, and sensitivity to market risk). Ratings for each of the six components and the final composite rating are on a 1 to 5 scale, with lower numbers indicating fewer or lesser regulatory concerns. Banks with a composite rating of 1 or 2 present few significant regulatory concerns and are considered healthy. In contrast, banks with ratings of 3 or greater present moderate to extreme levels of regulatory concerns and face much tougher supervision. In this paper, we refer to composite ratings of 3 or above as unsatisfactory ratings.

Aside from providing a summary measure of banking supervision that is easily comparable, CAMELS ratings are also relevant for important policy decisions. CAMELS determine insurance premiums on deposit insurance by the FDIC; access to credit from the Federal Reserve as the lender of last resort; licensing, branching and merger approvals; and eligibility for government programs such as the Troubled Asset Relief Program (TARP) and small business lending programs.

By the end of 2018, there were more than 50,000 financial examiners in the U.S., evenly split between state and federal regulators, earning a median salary of about $80,000. These examiners usually hold a bachelor's degree in accounting, finance, or economics and undergo at least one year of on-the-job training. On average, they have a tenure of around 14 years. Unlike private-sector counterparts, examiners' pay is not linked to bank performance. They are civil servants with high job security and fixed salaries that increase with seniority, along with some retention bonuses unrelated to bank outcomes.

II.C Examiner rotation and concurrent examinations

Banks in our main empirical sample are quasi-randomly assigned to examiners in the sense that banks are rotated across examiners *within* an agency, and rotated across agencies (state and federal) in alternate years. In this section, we describe how U.S. commercial banks are subject to CAMELS examinations by both state and federal regulators, the rationale for rotating individual lead examiners across banks, and why some banks receive concurrent exams from multiple agencies.

[10]Banks with assets below a minimum threshold are subject to exams every 18 months instead of every 12 months. This threshold has changed over time and since 2007 stands at $500 million for state member banks (SMBs) and non-member banks (NMBs). See the US Code Title 12, §1820 (d. 3) for an explicit codification.

By law, every insured depository institution must undergo a full-scope, on-site safety-and-soundness examination—commonly known as a CAMELS exam—on a regular cycle: typically every 12 months, or every 18 months for smaller institutions with strong ratings.[11] Commercial banks, which are the focus of our analysis, are generally supervised by both a state chartering authority and a federal regulator, such as the Federal Reserve, FDIC, or OCC. A key aspect of this interagency coordination is the use of alternating safety-and-soundness examinations between state and federal authorities.[12] The Riegle Community Development and Regulatory Improvement Act directed federal banking agencies to coordinate their exams in order to "minimize the disruptive effects" on bank operations.[13] In response, federal agencies issued joint guidance endorsing alternating supervision as the default approach for well-rated banks.

Regardless of which agency conducts the exam, regulators emphasize the importance of rotating individual examiners, especially the lead examiner (sometimes also called the "examiner-in-charge"). Examiner rotation promotes supervisory objectivity, reduces the risk of examiner "capture," and mitigates the potential for complacency or institutional bias.[14] By introducing fresh perspectives, examiner rotation enhances the credibility of supervisory assessments and prevents overly familiar relationships from forming between examiners and bank management. This policy is embedded in the supervisory frameworks of the OCC, FDIC, and Federal Reserve.[15]

When alternating supervision across state and federal agencies is deemed impractical due to a bank's complexity, risk profile, or the need for ongoing oversight by multiple regulators, concurrent examinations are conducted instead.[16] Federal agencies have established eligibility criteria for participation in the alternate-year examination program. Banks are excluded if they: have over $10 billion in assets; receive a composite CAMELS rating of 3 or higher; have unresolved compliance issues; or are undergoing significant events such as mergers, acquisitions, or leadership changes. Newly chartered (de novo) banks are also ineligible until they receive composite ratings of 1 or 2 in two consecutive examinations following their establishment.[17] In a concurrent exam, federal and state agencies perform coordinated on-site reviews during the same period but issue separate reports to fulfill their individual statutory responsibilities. Although examiners may share findings and co-

[11] FDIA, 1994, "Federal Deposit Insurance Act", 12, SC 1820d-6.

[12] FRS, 1995, "Interagency Statement on Guidelines for Relying on State Examinations" Board of Governors of the Federal Reserve System, SR 95-40.

[13] FDIA, 1994, "Federal Deposit Insurance Act", 12, SC 1820d-6.

[14] FHFA, 2017 "FHFA's Practice for Rotation of its Examiners Is Inconsistent between its Two Supervisory Divisions," Federal Housing Finance Agency Office of Inspector General.

[15] OCC, 2018, "Examination Strategy and Risk-Focused Examinations." Comptroller's Handbook.
FRB, 2024, "Approaches to Bank Supervision", Federal Reserve Board.
GAO, 2024, "Bank Supervision: Federal Reserve and FDIC Should Strengthen Supervisory Practices and Escalation Processes." U.S. Government Accountability Office.

[16] FRS, 1995, "Interagency Statement on Guidelines for Relying on State Examinations" Board of Governors of the Federal Reserve System, SR 95-40.
FFIEC, 1996, "Interagency Policy Statement on the Allowance for Loan and Lease Losses." 61 Federal Financial Institutions Examination Council, 61, 60642.

[17] OCC, 2008, "Bank Supervision Process" Comptroller's Handbook.
FRB, 2011, "Examination Strategy and Risk-Focused Examinations - Commercial Bank Examination Manual", Federal Reserve Board.

ordinate aspects of the review, each agency maintains full independence in its final supervisory assessments.

Our baseline empirical sample only includes banks that are subject to regular rotation across examiners within a single agency and across agencies (state and federal) in alternating years. We exclude banks subject to concurrent examinations, so we do not directly observe disagreement in the form of multiple lead examiners simultaneously assigning ratings to the same bank. However, as outlined in Section III, our empirical framework enables us to infer the extent of such disagreement. In a separate sample of concurrent examinations, we also directly measure disagreement.

II.D Data

We use a unique dataset, the National Information Center of the Federal Reserve System, covering the period from 1998 through the first quarter of 2020, of all on-site examinations of safety and soundness conducted by banking regulators. The data contain detailed financial information for depository institutions, regulated and select non-regulated institutions, as well as other institutions that have a regulatory or reporting relationship with the Federal Reserve System. The key data for the purposes of this study are unique bank identifiers, the lead examiner identity, the exam date, and the composite CAMELS rating and its component ratings. In contrast to several papers that have explored the determinants of supervisory ratings at the bank holding level (e.g., Berger et al., 2000), we employ the ratings at the level of the commercial bank, which is the entity level at which we observe the examiner rotations.

We merge this information with balance sheet measures of bank profitability and asset quality from Call Reports. Our main Call Reports variables are: Tier1 risk-based capital ratio, leverage ratio (Tier1 capital as a share of total risk-unweighted assets), return on assets, share of non-performing loans to total loans, and the delinquency rate of the loan portfolio. Delinquent loans include loans 30+ days past due and loans in nonaccrual status, and nonperforming loans include loans 90+ days delinquent and loans in nonaccrual status.

We restrict our sample to state non-member or state member banks, which are subject to regular examiner rotation. Within a region (a state for state regulators or a block of states for federal regulators), lead examiners and their teams are rotated across the set of banks located in each region. Our sample consists of banks subject to alternating examinations by state and federal regulators. Since states fall within larger federal regions, we consider the state as the relevant region for examiner rotation. Due to this regular rotation pattern, we assume that conditional on a state-year-quarter, the assignment of lead examiners to banks is uncorrelated with a bank's true quality (both observable and unobservable). Because assignment is not completely random, and considerations of examiner workload could potentially disrupt regular rotation, we also empirically support this assumption by showing that observable measures of bank quality are uncorrelated with examiner assignment, conditional on state-year-quarter fixed effects.

We further filter the sample by excluding exams where all components of the CAMELS rating are not scored or available. We exclude exams where identity of examiners is not tracked. We

also exclude targeted exams because of their exceptional nature relative to the routine safety and soundness examinations which are our focus. Finally, we exclude banks that do not display regular examiner rotation during our sample period (approximately 10 percent of the full sample). These banks appear to be depository institutions with peculiar purposes (e.g., Industrial Loan Companies (ILCs) or *de novo* banks). Since these banks do not satisfy our condition for identification that requires exogenous rotation of examiners, we exclude them from our sample.

While our data has the advantage of containing identifiers for the lead examiner in charge of each bank examination, the data does not contain examiner demographic information. Thus, we cannot study specific demographic questions such as gender differences in discretion. Instead, we focus on characterizing the total amount of heterogeneity in decision-making across examiners which can result in costly noise, and we show that variation in the weights these examiners put on specific issues is a major contributor to this heterogeneity.

III Framework for human discretion

III.A Discretion in cases with hard and soft information

We present a framework to clarify how discretion introduces both signal and noise into human decision-making. The goal is to distinguish between informative variation arising from soft information and unproductive variation arising from human noise. This structure helps identify the sources of disagreement across decision-makers and motivates empirical measures of discretion.

We consider a setting in which decision-makers $i \in \mathcal{I} = \{1, \ldots, I\}$ are randomly assigned to evaluate cases $j \in \mathcal{J} = \{1, \ldots, J\}$, each with an unobserved optimal decision Z_j^*. The actual decision rendered by decision-maker i on case j is Z_{ij}. We model human decision-making as

$$Z_{ij} = Z_j^* + b + e_{ij}, \tag{1}$$

where b captures a constant population-level bias (e.g., whether decisions are on average too tough or too lenient), and e_{ij} represents the additional error relative to the optimal outcome introduced by individual decision-makers. We assume $\mathbb{E}[e_{ij}] = 0$ and $\text{Var}(e_{ij}) = \sigma^2$, where σ^2 captures the extent of disagreement or noise across decision-makers who evaluate the same case. This formulation aligns with the notion that variation in decisions, conditional on the same information, reflects judgment inconsistency rather than legitimate heterogeneity in case content.

Let X_j represent hard information, i.e., a vector of observable characteristics for case j, corresponding to standardized, verifiable information. The expected decision based on hard information is $\mathbb{E}[Z_{ij} \mid X_j]$, which we estimate empirically as $\hat{Z}_{ij}$.

We define soft information as the deviation of the expected decision from this benchmark: $s_j = \mathbb{E}[Z_{ij}] - \mathbb{E}[Z_{ij} \mid X_j]$. Soft information represents information that is tied to case j that is only accessible through human judgment. Soft information includes subjective or contextual details—such as credibility, tone, or nuance—that are not captured by observable variables. We assume that

$\mathbb{E}[s_j] = 0$. Under these assumptions, a decision-maker with zero bias and error ($b = e = 0$) who observes both hard and soft information would arrive at the optimal decision Z_j^*.

We define discretion as the deviation between the actual decision and the benchmark predicted from hard information:

$$d_{ij} = Z_{ij} - \mathbb{E}[Z_{ij} \mid X_j] = s_j + e_{ij}. \tag{2}$$

This decomposition highlights a central idea: discretion can incorporate valuable soft information but also introduce error.

To further understand the nature of error, we decompose e_{ij} into two components: $e_{ij} = \mu_i + \epsilon_{ij}$, where μ_i reflects persistent tendencies of decision-maker i to be tough or lenient across cases and ϵ_{ij} captures residual, case-specific, inconsistency. We assume $\mathbb{E}[\mu_i] = 0$ with $\text{Var}(\mu_i) = \sigma_\mu^2$. We also assume $\mathbb{E}[\epsilon_{ij}] = \mathbb{E}[\epsilon_{ij} \mid i] = 0$, $\text{Var}(\epsilon_{ij}) = \sigma_\epsilon^2$, and $\text{Var}(\epsilon_{ij} \mid i) = \sigma_{\epsilon_i}^2$ which can vary across decision-makers.

To capture the magnitude of discretionary behavior, we define absolute discretion as the absolute value of the discretionary deviation:$|d_{ij}| = |s_j + \mu_i + \epsilon_{ij}|$. While d_{ij} captures the direction of deviation—whether a decision is more tough or lenient than predicted, $|d_{ij}|$ measures the magnitude of that deviation, irrespective of sign. For example, an examiner can have d_{ij} equal to zero on average across cases (no consistent bias toward leniency or toughness), while still exhibiting high absolute discretion if she heavily weighs gut feelings in either direction.

Note that the expected magnitude of absolute discretion, conditional on a decision-maker i, is $\mathbb{E}[|d_{ij,t}| \mid i] = \mathbb{E}[|s_{j,t} + \mu_i + \epsilon_{ij,t}|]$, where μ_i is a constant. For tractability, we assume the $s_{j,t}$ and $\epsilon_{ij,t}$ are distributed jointly normal with mean zero and covariance $\sigma_{s\epsilon_i}$. It is straightforward to show that $\mathbb{E}[|d_{ij,t}| \mid i]$ is increasing in $|\mu_i|$, $\sigma_{\epsilon_i}^2$, and $\sigma_{s\epsilon_i}$.[18] In other words, a decision-maker exercises greater expected absolute discretion if she has more extreme persistent error $|\mu_i|$, greater idiosyncratic inconsistency $\sigma_{\epsilon_i}^2$, or weighs the soft information of each case more heavily as reflected in $\sigma_{s\epsilon_i}$.

III.B Isolating error and measuring disagreement

We observe discretion d_{ij} but cannot separately observe soft information s_j and error $e_{ij} = \mu_i + \epsilon_{ij}$. To isolate the persistent component of error μ_i, we leverage the assumption that decision-makers are randomly assigned to cases. This implies that, in expectation, the distribution of soft signals is identical across decision-makers: $\mathbb{E}[s_j \mid i] = \mathbb{E}[s_j] = 0$. That is, while soft information varies across cases, each decision-maker is equally likely to receive high- or low-signal cases. Under this assumption, we can write: $\mathbb{E}[d_{ij} \mid i] = \mathbb{E}[e_{ij} \mid i] = \mu_i$, implying that the expected discretion for each decision-maker reflects only the decision-maker's persistent tendency to deviate from the

[18] Under these assumptions, $(|d_{ij,t}| \mid i)$ is a folded normal random variable, and $\mathbb{E}[|d_{ij,t}| \mid i] = \tau_i\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu_i^2}{2\tau_i^2}\right) + |\mu_i|\left[1 - 2\Phi\left(-\frac{|\mu_i|}{\tau_i}\right)\right]$, where $\tau_i = \sqrt{\sigma_s^2 + \sigma_{\epsilon_i}^2 + \sigma_{s\epsilon_i}}$.

benchmark. This expectation can be consistently estimated using decision-maker fixed effects.[19]

To estimate the dispersion in persistent error μ_i, we treat each decision-maker's average discretion as a noisy signal of their underlying type. For a decision-maker i with n_i cases, we compute $\bar{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} d_{ij}$. Because $\bar{\mu}_i$ may be imprecise due to limited sample size in a finite panel, we apply an Empirical Bayes (EB) shrinkage adjustment: $\bar{\mu}_i^{EB} = (1-\lambda)\bar{\mu}_i^{OLS} + \lambda\bar{\mu}$.[20] We measure persistent error dispersion as the variance of EB-adjusted decision-maker fixed effects $\hat{\sigma}_\mu^2 = \text{Var}(\bar{\mu}_i^{EB})$.

We define disagreement (also referred to as noise) as the variance of decisions across decision-makers evaluating the same case. Disagreement can be expressed as $\text{Var}(Z_{ij} \mid j) = \text{Var}(e_{ij}) = \sigma^2$, where $e_{ij} = \mu_i + \epsilon_{ij}$. Assuming $\epsilon_{ij} \perp \mu_i$, we can decompose disagreement as $\sigma^2 = \sigma_\mu^2 + \sigma_\epsilon^2$. This expression highlights two types of disagreement: persistent differences in decision-maker types and additional idiosyncratic inconsistency. $\hat{\sigma}_\mu^2$, as described above, offers an estimate of disagreement across decision-makers that arises from stable differences in their evaluation tendencies. It serves as a lower bound for the total amount of noise or disagreement due to human discretion, as this estimate excludes additional noise that could result from case-level inconsistency (ϵ_{ij}).

Note that, although we define disagreement as the variation in decisions for the *same* case, the above framework allows us to estimate the amount of disagreement even when we do not observe multiple decisions for the same case.

III.C Understanding the sources of disagreement

We define disagreement or noise as the variance of decisions across decision-makers evaluating the same case. In the previous section, we developed a conservative estimate of the amount of disagreement, $\hat{\sigma}_\mu^2$, using only variation in the persistent component of error, μ_i, across decision-makers. We now ask a more fundamental question: what drives disagreement? To answer this, we study the sources of variation in the observed decision. By decomposing $\text{Var}(Z_{ij} \mid j)$ into interpretable components, we can identify structural features—such as disagreement in interpreting component issues or in weighting decision components—that give rise to disagreement.

We begin by observing that most real-world decisions are not made as one-shot judgments, but instead by combining evaluations of multiple sub-issues. Let C_ℓ be the evaluation of component ℓ and w_ℓ be the weight placed on that component by the decision-maker. We model the final decision as an aggregation of judgments across multiple component issues:

$$Z_{ij} = \sum_{\ell=1}^{k} w_{\ell ij} C_{\ell ij}, \tag{3}$$

[19] In contrast, if random assignment fails, then $\mathbb{E}[s_j \mid i] \neq 0$, and we can no longer attribute average discretion to persistent error alone: $\mathbb{E}[d_{ij} \mid i] = \mathbb{E}[s_j \mid i] + \mu_i$.

[20] Empirical Bayes shrinkage adjusts each decision-maker's estimated fixed effect $\bar{\mu}_i^{OLS}$ toward the grand mean $\bar{\mu}$, based on the relative magnitude of signal and noise. The shrinkage weight λ is given by: $\lambda = \frac{\sigma_\nu^2/n_i}{\sigma_\mu^2 + \sigma_\nu^2/n_i}$, where σ_μ^2 denotes the variance of true decision-maker effects and σ_ν^2 the variance of idiosyncratic error. When n_i is small or the signal-to-noise ratio is low, more weight is placed on the overall mean.

where $C_{1ij}, \ldots, C_{kij}$ and $w_{1ij}, \ldots, w_{kij}$ are jointly independent. This formulation aligns with Information Integration Theory (see e.g., Anderson, 1971), which posits that people form judgments as weighted combinations of subjective inputs—a process observed across domains such as impression formation, moral reasoning, and preference judgments.

For brevity, we will omit the case subscript j in the remainder of this subsection—it should be understood that we are concerned with variation in decisions holding case j and its associated hard and soft information constant. We do not impose the constraint that the weights sum to one, both for tractability and because decision-makers, such as bank examiners in our setting, are not required to assign normalized weights across components. Allowing both C_ℓ and w_ℓ to vary across decision-makers i, we can express the variance in overall decisions as:

$$\text{Var}(Z_i) = \sum_{\ell=1}^{k} \mathbb{E}[w_{\ell i}^2] \cdot \text{Var}(C_{\ell i}) + \sum_{\ell=1}^{k} \text{Var}(w_{\ell i}) \cdot \mathbb{E}[C_{\ell i}]^2. \tag{4}$$

This decomposition reveals three contributors to disagreement in final decisions.

First, disagreement in the final decision increases with disagreement in component issues, $\text{Var}(C_{\ell i})$. Importantly, even issues that appear objective on their face can exhibit substantial variation in judgment. For example, consider a component-level decision $C_{\ell i} = \mathbb{E}[C_{\ell i} \mid X_\ell] + s_\ell + e_{\ell i}$, where $\mathbb{E}[C_{\ell i} \mid X_\ell]$ reflects the benchmark conditional on hard information, $s_{\ell i}$ captures soft signals, and $e_{\ell i}$ represents examiner-specific error. If $e_{\ell i} \approx 0$, then $\text{Var}(C_{\ell i}) \approx 0$. But if decision-makers inject either subjective interpretation or noise, we obtain $\text{Var}(C_{\ell i}) > 0$.

Second, disagreement increases when decision-makers assign high weight to highly subjective issues, i.e., those with large $\text{Var}(C_{\ell i})$, as seen in the first term of the variance decomposition.

Third, disagreement increases with heterogeneity in weights across decision-makers, $Var(w_{\ell i})$, particularly when greater variation in weights is applied to highly subjective issues (issues with large $E[C_{\ell i}]^2$). This is implied by the second term of the variance decomposition.

Together, the two terms of the variance decomposition imply that efforts to reach agreement on component issues may not lead to consensus on the final decision. Even if decision-makers agree on all component decisions ($\text{Var}(C_{\ell i}) = 0$), differing weights can still produce disagreement in final decisions. In addition, if most components are agreed upon but one highly subjective issue receives high weight, disagreement can persist.

III.D Real consequences of discretionary error

The previous section established that the error component of discretion contributes to disagreement, i.e., noise, in decisions. We now turn to an important applied question: does discretion matter for real outcomes? Specifically, we examine whether discretionary variation—particularly that driven by examiner-level error—has causal consequences for future bank performance. If discretion simply reflects benign differences in judgment that are neutralized elsewhere in the bank regulatory system, it may have little effect on bank outcomes. But, if discretion leads to variation in bank supervision that is not offset through other means, it may substantially impact bank behavior.

To evaluate this possibility, we test whether examiner discretion $d_{ij,t}$ affects future bank outcomes $Y_{j,t+1}$, conditional on observed covariates $X_{j,t}$:

$$Y_{j,t+1} = \alpha + \beta d_{ij,t} + \gamma X_{j,t} + \varepsilon_{j,t+1}.$$

For ease of interpretation, we let $Y_{j,t+1}$ represent bank risk, with higher values indicating greater safety and soundness concerns. Higher values of discretion $d_{ij,t}$ correspond to tougher supervisory decisions. We expect an exogenously (unrelated to bank fundamentals) tougher decision due to examiner discretion to cause the bank to engage in safer behavior, leading to a reduction in future measures of bank risk. Thus, we expect $\beta < 0$ as a measure of the causal effect of $d_{ij,t}$ on $Y_{j,t+1}$.

However, discretion may be endogenous. Examiner discretion includes bank soft information that is unobserved by the econometrician but predictive of future performance. We define soft information $s_{j,t}$ such that higher values correspond to great bank risk. If the measure of discretion is endogenous and includes $s_{j,t}$, the OLS estimate of β would be biased upward.

To isolate the component of discretion driven by examiner-level error μ_i, we construct a jackknife instrument using the leave-one-out average discretion across all other cases reviewed by examiner i:

$$\bar{d}_{i,-jt} = \frac{1}{N_i - 1} \sum_{(j',t') \neq (j,t)} d_{ij',t'}. \tag{5}$$

Under random assignment, and as the number of reviewed cases increases, this average converges in probability to the examiner's persistent error type: $\bar{d}_{i,-jt} \xrightarrow{p} \mu_i$. This instrument captures examiner-level tendencies that are orthogonal to case-specific variation, excluding the soft signal $s_{j,t}$ and idiosyncratic noise.

We estimate a two-stage least squares (2SLS) model.

$$\text{First stage:} \quad d_{ij,t} = \pi \bar{d}_{i,-jt} + \eta X_{j,t} + \xi_{j,t}, \tag{6}$$

$$\text{Second stage:} \quad Z_{j,t+1} = \alpha + \beta^{IV} \hat{d}_{ij,t} + \gamma X_{j,t} + \varepsilon_{j,t+1} \tag{7}$$

A significant $\hat{\pi}$ in the first stage estimation is evidence that decision-makers have a persistent component in their errors. A significant $\hat{\beta}^{IV}$ estimate indicates that persistent examiner tendencies—those unrelated to the hard and soft information content of the case—have real consequences for bank outcomes.

Using the 2SLS approach above, we will show that examiner discretion causally distorts bank behavior. In what follows, we provide an illustration of how real distortions in bank behavior due to examiner discretion could lead to *costly* noise. Because the precise cost depends on the social planner's objective function, this example is meant to be suggestive rather than definitive.

We assume that an optimal composite rating exists for each bank examination. One way to think about this is that these ratings translate into optimal outcomes in terms of bank capitalization

and lending. If examiner discretion moves the composite CAMELS rating away from its optimal level, it will also move capitalization and lending away from the optimum, creating costly noise. Consider a case j which has an optimal outcome Z_j^* (for brevity, we omit the j subscript in the remainder of this subsection). As econometricians, we do not observe Z^* and we make no assumptions regarding b, the extent to which the population of examiners is too tough or too lenient in expectation.

Let $F(Z - Z^*)$ represent the cost of deviations from the optimal case outcome. We follow Kahneman, Sibony, and Sunstein (2021) and assume a quadratic cost function. This allows us to decompose the cost of discretion into separate bias and noise terms.

Suppose $F\left(Z - Z^*\right) = (Z - Z^*)^2$. The expected cost of discretion is:

$$E\left[(Z - Z^*)^2\right] = E\left[(b + e)^2\right] = b^2 + \sigma^2. \tag{8}$$

b^2 represents the cost due to population-level bias (i.e., examiners on average being too lenient or tough) and σ^2 represents the cost due to noise.[21] The separability of the two terms implies that, even if decision-makers are unbiased ($b = 0$), noise would still have a cost of σ^2. The model further implies that two social planners who disagree about whether bank examiners are too tough or lenient on average (i.e., whether $b > 0$ or $b < 0$), would nevertheless agree that the cost of noise is equal to σ^2. Further, for equal amounts of bias and noise, the marginal cost of an additional unit of noise is the same as the marginal cost of an additional unit of bias, implying that the problem of noise is as important as the problem of bias.

Finally, our paper main focus is on identifying the determinants of discretion and quantifying its *ex post* consequences and trade-offs. Accordingly, the simple framework we present captures only the *ex post* effects of discretion. However, the anticipation of discretion by decision-makers may also shape the *ex ante* behavior of affected entities. We will present suggestive evidence that banks in states where examiners exercise greater discretion decisions take conservative actions *ex ante* by maintaining higher capitalization and lower lending. This anticipation effect could be costly or beneficial depending on the social planner's preferences.[22]

III.E The value of discretion

While the previous section focused on how discretion could distort outcomes and lead to costly error, discretion is not inherently detrimental. It may also capture soft information that

[21] The quadratic loss function is equivalent to the Mean Squared Error (MSE) function, which is very commonly used in statistics, econometrics, and machine learning to measure the quality of an estimator. The quadratic form offers a convenient approximation for settings in which the cost of the deviation from an optimum is convex with respect to the absolute magnitude of the deviation.

[22] The question of the optimal degree of discretion, including *ex ante* consequences, is outside the scope of our paper. Theoretical insights into this question can be found in Leitner and Williams (2022), which models the related question of the optimal amount of model secrecy in bank stress tests. Noise in examiner decision-making could have a similar effect to model secrecy because both contribute to regulatory uncertainty. Extending the conclusions of the Leitner and Williams model to our setting implies that optimal noise is single-peaked. Thus, regulators may wish to carefully consider the optimal degree of discretion along with the optimal degree of model secrecy.

is economically meaningful but unobservable in structured data. In this section, we turn to the potential benefits of discretion, asking whether discretionary variation contains predictive signal about future outcomes.

We revisit the predictive regression introduced earlier, now shifting our focus from identifying distortions due to examiner error to assessing the value of discretion:

$$Y_{j,t+1} = \alpha + \beta^{OLS} d_{ij,t} + \gamma X_{j,t} + \varepsilon_{j,t+1}, \tag{9}$$

where $Y_{j,t+1}$ denotes bank risk in the next period, and $X_{j,t}$ includes hard information available at time t.

$\beta^{OLS} = \frac{\text{Cov}(d_{ij,t}, Y_{j,t+1})}{\text{Var}(d_{ij,t})}$ measures the predictive relation between discretion and future bank risk. Given that discretion reflects both soft information and error ($d_{ij,t} = s_{j,t} + e_{ij,t}$), β^{OLS} captures a combination of forces related to signal and noise. Soft information about bank risk should be *positively* associated with future bank risk $Y_{j,t+1}$ whereas exogenously tougher bank supervision as reflected in $e_{ij,t}$ causes a reduction in future bank risk. This latter channel is captured by $\beta^{IV} < 0$ in our 2SLS estimation.[23]

Specifically, we can decompose the relation between discretion and future bank risk as follows:

$$\text{Cov}(d_{ij,t}, Y_{j,t+1}) = \text{Cov}(s_{j,t}, Y_{j,t+1}) + \text{Cov}(e_i, Y_{j,t+1}), \tag{10}$$

and rearrange terms to obtain an empirical proxy:

$$\beta^{OLS} \approx \frac{\text{Cov}(s_{j,t}, Y_{j,t+1})}{\text{Var}(d_{ij,t})} + \beta^{IV} \cdot \frac{\text{Var}(e_i)}{\text{Var}(d_{ij,t})}. \tag{11}$$

Given that $\beta^{IV} < 0$, a finding of $\beta^{OLS} > 0$ would imply that the soft information effect dominates, and that discretion on net adds to predictive power.

While discretion can improve decision quality by incorporating soft information, more discretion is not always beneficial. In fact, when variation in discretion reflects examiner-specific inconsistency rather than insight, it may dilute the predictive value of decisions. Below, we investigate how the informativeness of discretion varies with its intensity, and whether greater use of discretion necessarily leads to better outcomes.

We use absolute discretion, $|d_{ij,t}|$, as our measure of the intensity of discretion. We are interested in whether examiners who exercise greater discretion, i.e., those with greater $\mathbb{E}[|d_{ij,t}| \mid i]$, make decisions that are more predictive of future bank risk. As shown previously, a decision-maker exercises greater expected absolute discretion if she has more extreme persistent error $|\mu_i|$, greater idiosyncratic inconsistency $\sigma^2_{\epsilon_i}$, or weighs soft information more heavily $\sigma_{s\epsilon_i}$. Whether more discretion leads to decisions that are more predictive of future bank risk is theoretically ambiguous. We investigate the question empirically by testing how the predictive coefficient β^{OLS} varies with the examiner's absolute discretion.

[23] While we empirically estimated β^{IV} using variation in μ_i, β^{IV} captures the causal effect of a general exogenous change in supervision toughness on bank outcomes. Thus, we also assume that $\beta^{IV} = \frac{\text{Cov}(e_{ij,t}, Y_{j,t+1})}{\text{Var}(e_{ij,t})}$.

In summary, discretion adds value because it allows the incorporation of soft signals, but discretion can also reflect noise which can potentially weaken the predictive relation between decisions and future bank risk. Our framework highlights that more discretion does not necessarily imply better decisions.

IV Results

IV.A Summary statistics

We begin by presenting summary statistics of our baseline data in Table 1. Our data covers the period from 1998 to the first quarter of 2020. Panel A shows that we observe 2,407 lead examiners and 14,679 examinations in our cleaned data sample.[24] Panel B shows the distribution of the composite and component ratings. Most banks receive a composite rating of 1 or 2 which is considered safe, and 15% receive an unsatisfactory composite rating of 3, 4, or 5, which would lead to additional bank oversight and regulation. Panel C summarizes the transition probabilities between the bank's current composite rating and the bank's next composite rating in the following year. We see that the majority of banks retain their current composite rating in their next exam. However, transition probabilities toward different ratings are non-trivial. For example, a healthy bank with a current composite rating of 2 faces a 7.8 percent probability of being rated as unsatisfactory in the next exam (3 or higher).

Panel D summarizes bank observables as of the quarter end immediately before the bank examination is completed and the rating is released. We adopt this timing convention to ensure that the bank variables used in our analysis are observable to the examiner at the time of the examination. *Tier1 ratio* is defined as tier1 capital / risk-adjusted assets. *Loan growth* is defined as log(loans) – log(loans one year ago). *Leverage* is defined as tier1 capital / assets. *ROA* is defined as net income / assets. *NPL ratio* is defined as loans >90 days past due + non-accrual loans / total loans. *Delinq ratio* is defined as delinquent loans / loans. *Efficiency* is defined as non-interest expenses / revenue. All ratios are computed as percents with numerators and denominators measured in thousands of US dollars. Bank size within our sample is right-skewed with mean assets of 2.3 billion US$ and median assets of 186 million US$.

Note that because we limit our sample to banks subject to regular examiner rotation, we exclude the largest banks in the US economy. Large regional banks are included in our sample, such as Silicon Valley Bank and First Republic Bank with assets of several hundred billion dollars. As evidenced by the collapse of Silicon Valley Bank and First Republic Bank in 2023, the health of these regional banks can have a large impact on the broader financial system. We further note that even the largest banks in the US are still subject to CAMELS examinations, so examiner discretion

[24] Lead examiners within our sample are associated with an average of six exams. We observe relatively few regular exams per lead examiner due to three reasons: (1) most examiners serve for long periods as non-lead examiners before rising to the position of lead examiner, (2) lead examiners also work on non-standard exams for problem banks and these examinations are excluded from our sample because they fall outside the regular rotation schedule, and (3) examiners have a relatively high rate of rate turnover.

could, in principle, also have a major impact on very large banks. However, without a regular rotational structure, we are unable to credibly estimate the exact impact of examiner discretion on these very large banks.

Figure 1 shows the average CAMELS rating across all banks in our sample over time. We find that the average composite rating and component ratings remained approximately constant from 1998 to the mid 2000's, rose dramatically in the years leading up to the Financial Crisis, peaked in 2009, and then declined to pre-crisis levels by 2013.

IV.B Measuring discretion

We begin by estimating the predicted CAMELS composite rating conditional on bank observables. Using our full data sample, we estimate the following regression:

$$Rating_{ijrt} = \beta X_{it} + \gamma_{rt} + \varepsilon_{ijrt} \tag{12}$$

Observations are at the exam level, and is indexed for examiner i, bank j, state r, and year-quarter t. $Rating_{ijrt}$ is the composite CAMELS rating, X_{it} represents bank observable characteristics described in Panel D of Table 1, γ_{rt} represents state-year-quarter fixed effects, and ε_{ijrt} represents the residual error term, which we allow to be clustered at the examiner level. Regression results are presented in Column 1 of Table 2.

We define $RatingPred_{ijrt}$ as the predicted value from this regression, representing the expected rating based on hard bank observables and region-time trends. As discussed in Section III, $RatingPred_{ijrt}$ does not represent the correct or optimal decision. Nonetheless, it provides a useful benchmark: under random assignment, systematic deviations from $RatingPred_{ijrt}$ can be used to measure each examiner's persistent error μ_i.

For each observation, we define $Directional\_Discretion_{ijrt} = Rating_{ijrt} - RatingPred_{ijrt.}$. A positive value implies the examiner gave a tougher rating than predicted based on observables. We also define $Absolute\_Discretion_{ijrt} = |Rating_{ijrt} - RatingPred_{ijrt}|$. Zero $Absolute\_Discretion_{ijrt}$ implies the examiner gave a rating exactly as predicted by the observable bank characteristics, and more positive values imply the examiner exercised more discretion in either the tougher or more lenient directions.

IV.B.1 Examiner-level discretion

We aggregate to the examiner level by taking the average of each measure across all exams conducted by each examiner to form $Examiner\_Directional\_Discretion_i$ and $Examiner\_Absolute\_Discretion_i$. An examiner with high $Examiner\_Absolute\_Discretion_i$ is a person who tends to deviate from the predicted average rating in either direction while an examiner with high $Examiner\_Directional\_Discretion_i$ is a person who issues tougher ratings on average, conditional on observables. It is possible for an examiner to have zero directional discretion but high absolute discretion. Such an examiner would not be more lenient or tough than other examin-

ers on average, but may more heavily weight soft information or gut feelings in either direction. In later analysis, we use a leave-one-out version of $Examiner\_Directional\_Discretion_i$ to estimate the causal effect of exogenous variation in ratings on bank outcomes.

We also account for the possibility that our measures of discretion may be distorted due to integer rounding. Examiners are constrained to assign integer ratings 1 through 5 for the composite and component ratings. In contrast, the predicted rating from regression (4) can be any real value. Since we measure directional discretion and absolute discretion as ($Rating_{ijrt}$ – $RatingPred_{ijrt}$) and $|Rating_{ijrt}$ – $RatingPred_{ijrt}|$, respectively, we may estimate non-zero discretion for examiners who actually exercise zero discretion. Integer rounding does not introduce a bias to the regression results presented in the remainder of the paper, because the regression analysis compares differences across examiners in the same region-quarter, who are exposed to the same set of banks. However, integer rounding does matter for the interpretation of the magnitude of discretion. To address this issue, we round the predicted rating from regression (4) to the nearest integer value, 1 through 5, and measure directional and absolute discretion relative to the integer-rounded version of predicted rating.

IV.B.2 Examiner rotation

Before interpreting the magnitudes of the examiner-level measures of discretion, we first present empirical evidence in support of our identifying assumption that examiners are assigned to banks within a state-year-quarter in a way that is uncorrelated with the bank's true quality. One potential concern is that, when a bank becomes troubled, the bank may be more likely to be assigned to lead examiners who have more experience examining troubled banks. If so, we should find that examiners with higher leave-out-mean ratings (these are examiners who have more experience examining other troubled banks and/or are tougher) are more likely to be assigned to banks with weaker observable measures of bank quality. We show that examiner leave-out-mean ratings are uncorrelated with observable measures of bank quality within a location-time period. Thus, the data is consistent with the view that examiner assignment is uncorrelated with bank quality.

It remains possible that, holding constant observable measures of bank quality, unobservably worse banks are assigned to examiners with greater or lesser degrees of directional and absolute discretion. To address this concern, we proxy for unobservable (at the time of the examination) bank quality with the bank's future change in performance of its existing loan portfolio, as measured by the change in the bank's non-performing loan and delinquency ratios in the quarter after the rating is released relative to the quarter before the rating is released. We find that examiner leave-out-mean discretion is uncorrelated with these proxies for unobservable bank quality.

Specifically, Table 3 tests whether bank characteristics at the time of the exam and future changes in loan performance are correlated with examiner directional and absolute discretion. If examiners are rotated across banks within regions on a regular schedule, these variables should be uncorrelated with measures of examiner directional and absolute discretion, holding the region and time period fixed. Panel A regresses bank characteristics at the time of the exam and future changes in loan performance on the examiner's leave-out-mean directional discretion, as defined previously.

Panel B presents the same regressions, with the independent variable as the examiner's leave-out-mean absolute discretion, defined as the examiner's average absolute discretion across other exams, excluding the current observation. In all cases, bank observables and future changes in loan performance are not significantly correlated with examiners' directional and absolute discretion.

In Panel C, observations are ordered by bank-time, and we regress the leave-out-mean directional discretion and absolute discretion of the current examiner on the leave-out-mean directional discretion and absolute discretion of the examiner who previously examined the bank. In both cases, we find that the discretion of the current examiner is not related to the discretion of the previous examiner who was assigned to the bank. This shows that there is no serial correlation in assignment of a given bank to tougher or more discretionary examiners.

IV.B.3 Magnitudes of examiner-level discretion and disagreement

Building on the above evidence, we assume that examiners are assigned to banks within a state-year-quarter in a way that is uncorrelated with the bank's true quality. This random rotation enables us to interpret examiner-level average discretion. As explained in our framework in Section III, if examiners are matched to banks with similar expected values of hard and soft information within a region and time period, systematic differences in their rating behavior reveal their persistent error μ_i which we can use to estimate disagreement, defined as the variance in decisions if multiple examiners were to simultaneously examine the same bank.

Panels A and B of Table 4 summarize directional discretion and absolute discretion at the exam and examiner levels, respectively. The full distributions are plotted in Figure 4. We find that directional discretion varies widely across examiners. A standard deviation in examiner-level directional discretion is 0.18 points, or 0.19 after adjusting for integer rounding. As described in Section III, these estimates of the variance of examiner-level directional biases will be biased upwards due to measurement error in finite panels. After applying an Empirical Bayes shrinkage adjustment, we estimate that a standard deviation in examiner-level directional discretion is $\hat{\sigma}_\mu = 0.149$. Detailed output for the shrinkage adjustment is provided in the Appendix. As shown in Section III, $\hat{\sigma}_\mu$ represents a conservative estimate of disagreement, i.e., the variation in decisions if multiple examiners were to simultaneously examine the same bank.

We also find a substantial degree of examiner absolute discretion. The average examiner deviates from the predicted rating in either direction by 0.26, or 0.15 after adjusting for integer rounding. There is also substantial dispersion in the exercise of absolute discretion across examiners. We find that the standard deviation in absolute discretion across examiners is 0.24, or 0.36 after adjusting for integer rounding. This indicates that some examiners exercise significant discretion in either direction while others choose ratings that are close to the predicted values conditional on bank observables.

Next, we show that these measures of disagreement are economically large and meaningful. We show that the majority of instances in which banks experience a change in their CAMELS rating (relative to their rating in the previous year) is due to changes in examiner assignment rather than

changes in true bank quality. Using simulations described in detail in the Appendix, we estimate the percentage of banks that are assigned a higher or lower rating purely due to examiner discretion. We find that the distribution of examiner discretion implies that healthy banks that would otherwise receive a rating of 2, which compose the majority of our sample, are exposed to a 4.2% probability per exam of being rated an unsatisfactory 3 or higher. Likewise, 5.0% of banks that would have gotten a rating of 2 absent discretion receive a rating of 1 due to discretion. These magnitudes are large compared to the overall transition probability of 6.5% and 8.6% that a bank moves from a rating of 2 to a rating of 3 and 1, respectively, as shown in Table 1 Panel C. Thus, a majority of cases in which banks receive a different rating than in the previous year could be due to changes in examiner assignment rather than changes in true bank quality.

Another way to evaluate the magnitudes of discretion is to relate our estimates to research by Agarwal et al. (2014), who showed that state bank examiners are predictably more lenient than federal examiners, possibly due to incentive differences across the two government institutions. In Appendix Table 1 Panels A through D, we present the distribution of directional and absolute discretion separately for bank examinations conducted by federal and state examiners. We find that agency differences indeed matter, but account for only a small fraction of the overall variation in discretion across examiners. Consistent with Agarwal et al. (2014), federal examiners in our sample assign higher ratings conditional on bank observables: federal examiners' directional discretion are on average 0.07 points higher than that for state examiners. However, within the samples of only federal or only state examiners, we see substantial variation in discretion across examiners, approximately equal in magnitude to the total variation across examiners in the full sample. After applying an Empirical Bayes shrinkage adjustment, we find that the *within-agency* standard deviation in examiner-level directional discretion $\hat{\sigma}_{\mu}$ exceeds 0.14 for both state and federal examiners. Thus, the variation in persistent error due to discretion within agencies exceeds the variation across agencies.

In supplementary analysis reported in Appendix Table 4, we explore how the magnitudes of directional and absolute discretion vary with examiner experience. We find that more experienced examiners apply greater absolute discretion and are tougher (higher directional discretion) than less experienced examiners. Overall, we find no evidence that examiners exercise less discretion as they gain experience.

A potential concern with the measure of discretion presented in this section is that examiners' true model of bank health may be a non-linear function of observable bank hard information plus discretion. We can test whether our estimates of discretion are robust to using a much more flexible specification to control for the relation between ratings and bank observable variables X_{it}. In the Appendix, we control for linear and quadratic terms for each observable bank characteristics, as well as all two-way interactions between bank characteristics, following the approach in Iyer et al. (2016). We find that our measures of the quantity of discretion (Appendix Table 2) and the consequences of discretion (Appendix Table 3, as discussed in the next section) remain similar in magnitude to those in our baseline tests. These supplementary results suggest that our measures of discretion are

not sensitive to model misspecification of bank hard information observables.

IV.C Direct measure of disagreement

In our baseline empirical sample, bank examinations are not conducted concurrently, so we cannot directly observe examiner disagreement. Instead, we estimate disagreement indirectly by leveraging the quasi-random rotation of examiners across banks within a region. As discussed in Section III, because examiners are assigned to banks with similar expected levels of hard and soft information within a given region and time period, systematic differences in their rating behavior reveal disagreement.

For a smaller data sample of 410 observations (not included in our baseline sample), we observe examiner disagreement directly. In these cases, multiple examiners from different agencies assess the same bank in the same quarter. Table 5, Panel A, reports the probability of disagreement on composite and component ratings during these concurrent exams. Disagreement is common: examiners assign different composite ratings in 28% of cases. There is also substantial disagreement across all component ratings, with variation in disagreement rates by component. The highest rate of disagreement occurs in the management rating (31%), followed by the asset rating (23%). The remaining components—capital, earnings, liquidity, and sensitivity—each exhibit disagreement rates below 20%. Importantly, the disagreement measured here does not reflect minor differences in opinion. Because ratings must take integer values, any disagreement implies that two lead examiners assigned ratings that differ by at least one full point.

Panel B of Table 5 shows that banks subject to concurrent exams are broadly similar to those in our main sample based on observable indicators of bank health, with one key exception: they exhibit significantly higher non-performing loan ratios. This pattern aligns with the policy described in Section II.C, which allows banks flagged for potential risk to be removed from the regular alternating-agency rotation and instead examined concurrently by multiple agencies. We do not claim that concurrent exams are randomly assigned—indeed, this non-randomness is why we exclude them from our main estimation sample, which depends on quasi-random examiner rotation. Nonetheless, the concurrent exam data offer valuable insights into which component issues disagreement is most likely to arise. The results suggest that ratings of management and assets are relatively more subjective—that is, more prone to disagreement—although disagreement is present across all rating components.

IV.D Why do examiners disagree? Sources of disagreement

So far, we have shown that examiners discretion is associated with substantial disagreement. However, bank examiners are experienced professionals who are trained to assess bank safety and soundness. Why would two examiners, upon observing the same bank information, decide on different ratings? In this section, we show the important role of weights in disagreement.

As discussed in Section III.C, the final decision can be modeled as a weighted sum of judgments on individual component issues. Disagreement in the final decision can arise through three main

channels. First, it increases with the level of disagreement across the component issues themselves. Second, it grows when examiners place greater weight on component issues that are more prone to disagreement. Third, disagreement rises with heterogeneity in the weights assigned by different examiners—especially when this variation in weights is associated with component issues that are highly subjective.

IV.D.1 Disagreement in component issues

Regarding the first channel, where disagreement in the final decision increases with disagreement in component issues, we have already documented empirical evidence of disagreement in all components in Table 5. The most subjective components are management quality and asset quality. Notably, no component is entirely objective—even liquidity, the component with the highest agreement, has a disagreement rate exceeding 0.13.

As further evidence of disagreement in component ratings, we return to our baseline empirical sample—which includes banks subject to regular examiner rotation—and show that examiners' persistent biases significantly predict their decisions for each component issue. Specifically, for each component rating l and examination, we compute $Directional\_Discretion_{ijrtl} = Rating_{ijrtl} - RatingPred_{ijrtl}$, defined as the actual component rating minus the predicted rating based on observable hard information. We then estimate Equation 6 separately for each component rating—that is, we regress the current component rating on the examiner's average discretion for that component (calculated excluding the current exam), controlling for bank observables and state-year-quarter fixed effects. A significantly positive coefficient on the examiner's leave-out-mean direction discretional indicates that an examiner has a persistent tendency toward leniency or toughness in assessing that component.

We find positive and significant coefficients across all components in Table 6. For example, examiners who are consistently tougher than predicted in assessing the capital adequacy component in other exams are also tougher than predicted in the current exam. This pattern implies that disagreement extends to each component, as examiners' persistent biases systematically influence each component rating.

IV.D.2 Greater weight on more subjective components

Regarding the second channel, where disagreement in the final decision increases when greater weight is placed on component issues with greater disagreement, we find that examiners place the greatest weight on the most subjective component: management quality.

Panel A of Table 7 shows how the composite rating is related to the component ratings. Column 1 presents an unconstrained OLS regression of composite ratings on component ratings. Column 2 estimates the same regression subject to the constraint that the coefficients sum to one. The coefficients in these regressions can be thought of as weights on the component ratings, with or without the restriction of normalization of weights.[25] Notably, these estimates of weights do not

[25]These regression coefficients estimate the average weights on component ratings if we assume that the weights

require any assumptions about the randomness of examiner assignment to banks or measurement of bank observables.

We find that examiners, on average, place the highest weight, 48%, on the Management quality component rating, which as shown earlier, exhibits the highest level of disagreement. The second-highest weight, 16%, is assigned to Asset quality, which also has the second-highest level of disagreement. In contrast, examiners assign considerably less weight to the Capital, Earnings, Sensitivity, and Liquidity components. These patterns are illustrated in Figure 2, which presents the graphical counterpart to the regression results in Column 2 of Table 7.

We also find that examiners place substantial weight on assessments of management quality across a variety of subsamples. In Panel B of Table 7, we assess how weights on component ratings have changed over time. We divide the sample into three periods relative to the Great Recession: the pre-recession, recession, and post- recession periods. We find that examiners have always heavily weighted assessments of management quality, and the weight on management has grown to 54% in the post-crisis period.

Finally, we ensure that our finding of high weight placed on subjective component issues is robust to a more flexible model that accounts for the categorical nature of composite CAMELS ratings, which are restricted to integer values from 1 to 5. In the previous specification, we assumed equal spacing between ordinal values—for example, that the difference between ratings of 1 and 2 is equivalent to the difference between 4 and 5. In Column 3 of Table 7 Panel A, we estimate an ordered logit model, treating the composite CAMELS rating as an ordinal dependent variable. In this model, the numeric values of the outcome are irrelevant except that higher values indicate greater perceived bank risk. The reported coefficients reflect the log-odds impact of each component issue on the probability of receiving a higher composite rating category. We find that the relative magnitudes of the component coefficients in the ordered logit model (Column 3) closely mirror those in the OLS model (Column 1). Management ratings continue to receive the greatest weight, followed by Asset ratings.

The high weight placed on the management component is especially remarkable in that assessment of management quality is necessarily subjective. According to FDIC guidance, examiners should judge management quality based on "the level and quality of oversight and support by the board and management, ... the ability of the board and management to plan for, and respond to, risks, ... the extent of dominant influence or concentration of authority, ... [and management's] demonstrated willingness to serve the legitimate banking needs of the community."

These results highlight a key reason why examiners disagree in their final decisions: they assign relatively greater weight to components for which there is more disagreement, in particular the Management component rating. The heavy emphasis on perceptions of management quality aligns

are homogenous (constant) across examiners or if the weights are heterogeneous across examiners but uncorrelated with the component ratings (this would be the case if, e.g., some examiners always weight one component more heavily than others because they believe the component is more important). In the case in which assigned weights are correlated with assigned component ratings (e.g., examiners weight one component more when that component has a higher rating), the regression will not recover average weights. Rather, the regression coefficients estimate the marginal impact of each component rating on the composite rating, holding the other component ratings constant.

with a well-documented form of discretion in the psychology literature, where decision-makers tend to overweight inferences drawn from face-to-face interactions. For example, Levine, McCornack, and Park (1999) find that individuals often assign excessive importance to cues such as facial expressions, language, and interpersonal interactions.

IV.D.3 Heterogeneity in weights

Finally, we demonstrate that disagreement in final decisions arises because examiners disagree on the weights they assign to each component. Even when there is complete agreement on component assessments, differing weights can still lead to divergent final decisions. As discussed in Section III.C, such disagreement intensifies when greater variability in weights is applied to highly subjective components. Our analysis reveals substantial heterogeneity in weights across examiners, particularly for the Management component rating—the most subjective of all components.

To estimate heterogeneity in weights across examiners, we restrict our baseline sample to those associated with at least 10 exams, yielding 415 unique lead examiners. For each examiner, we regress the composite rating on the component ratings to recover examiner-specific weights. Table 8 reports the average and standard deviation of the weights assigned to each component across examiners. Once again, we find that examiners place the greatest weight—approximately 48%—on management quality. We also observe substantial heterogeneity in weights across examiners: the standard deviation exceeds 11% for all components and is highest, at 17%, for the management component—the same component that exhibits the most disagreement.

Next, we explore how the weighting function over component ratings varies with directional discretion and absolute discretion. The results are reported in Table 9, with a graphical representation in Figure 3. Column 1 restricts the sample to observations corresponding to observations in the top and bottom quartile of directional discretion, with quartile 4 representing tougher examiners (those who assign positive residual ratings). We regress the composite rating on the component ratings and the interaction between the component ratings and an indicator for quartile 4. The coefficients on the component ratings represent the weights for each component for observations in quartile 1. The coefficients on the interaction terms represent the difference in weights for quartile 4 relative to quartile 1. Column 2 estimates the same regression, with quartiles sorted by absolute discretion, with quartile 4 representing greater absolute discretion. Note that directional discretion and absolute discretion have a correlation of approximately zero in our sample, so it is worthwhile to separately examine how weights vary with these two measures of discretion.

We find that high positive directional discretion (quartile 4) is associated with significantly greater weight placed on Management (an increase of 25% relative to quartile 1), Liquidity (52%), and Sensitivity (102 %), and lower weight placed on Earnings (a decrease of 14%). Greater absolute discretion corresponds to significantly higher weight on Assets (13%) and Management (19%), and lower weight on Capital (16%), Liquidity (45%), and Sensitivity (26%). These patterns suggest that examiner disagreement in weights varies systematically with both directional and absolute discretion. In particular, examiners who exercise more discretion—i.e., those with higher absolute

discretion—place greater weight on the two most subjective components (Management and Assets) and less weight on the less subjective components, compared to examiners who exercise less discretion.

IV.E Real consequences of discretion

Examiner discretion may lead to large variation in CAMELS ratings, holding bank fundamentals constant. As discussed before, these ratings are central to how banks are regulated in the US. In this section, we assess how an exogenous change in CAMELS ratings induced by examiner discretion affects bank behavior.

Following the framekwork in Section III.D, we exploit the fact that examiners are rotated across the banks within each region and assume that assignment of banks to a tougher or more lenient examiner is uncorrelated with true bank quality, within a state x year-quarter. To measure each examiner's tendency to be tough or lenient, we create a measure of each examiner's leave-out-mean directional discretion, $Directional\_Discretion\_LO_{i,\text{-}jt}$, equal to the average directional discretion across all exams for examiner i, excluding the current exam. This approach is akin to a jack-knife measure of each examiner's persistent error μ_i, where we exclude the current observation so that unobserved bank quality for the current examination does not affect our measure.

We then estimate the following jack-knife instrumental variables strategy:

$$Rating_{ijt} = \pi Directional\_Discretion\_LO_{i,-jt} + X_{jt} + \gamma_{rt} + \varepsilon_{ijrt} \tag{13}$$

$$BankOutcome_{j,t+1} = \beta^{IV} \widehat{Rating}_{ijt} + X_{jt} + \gamma_{rt} + \eta_{ijrt} \tag{14}$$

We use $Directional\_Discretion\_LO_{j,\text{-}it}$ as the excluded instrument, and measure bank outcomes four quarters after the quarter of the current examination. Note that, while we use variation in examiner's persistent error component μ_i to estimate the causal effect of CAMELS ratings, this causal magnitude applies broadly, including to ratings that vary due to general examiner error $e_{ij} = \mu_i + \epsilon_{ij}$. We also note that our instrumental variables procedure will capture the overall effect of a tougher examiner and rating, including examiner guidance to management in addition to the direct effect of a higher rating.[26]

Column 1 of Table 2 shows the relation between the composite rating of the current exam and bank observables. Column 2 presents the first stage regression in the instrumental variables estimation, with the examiner leave-out-mean directional discretion as the excluded instrument. We find that $Directional\_Discretion\_LO_{j,\text{-}it}$ strongly predicts the composite rating, conditional on bank observables. A one-unit increase in $Directional\_Discretion\_LO_{j,\text{-}it}$ is associated with a 0.17 unit higher rating for the current exam. These results show that, while we observe a short panel per lead examiner, we are able to estimate a significant first stage in which the examiner's leave-

[26] In Appendix Figure A1, we follow Angrist and Imbens (1994) and show that the cumulative distribution function of the CAMELS ratings decision for high values of the instrument stochastically dominates the CDF of the CAMELS ratings decision for low values of the instrument. This supports the monotonicity assumption underlying our IV strategy, although we acknowledge that we cannot directly test the monotonicity assumption.

out-mean strongly predicts her current rating. The F-statistic on the excluded instruments in the first stage is well above 10, the rule of thumb threshold for weak instruments.

Panel A of Table 10 presents our estimates of the causal effect of higher composite ratings on future bank outcomes. Bank outcomes are measured four quarters after the quarter in which the current exam rating is finalized, or as of the next exam in the case of Column 1 which examines the next rating. The reported estimates are from an instrumental variables (2SLS) estimation in which the composite rating is instrumented with the examiner's leave-out-mean directional discretion. We find that an exogenously higher rating leads to changes in banks outcomes that correspond with the bank becoming less risky and more sound. In other words, we estimate that $\hat{\beta}^{IV} < 0$, as defined in Section III.D.

An exogenous one-point increase in ratings for the current exam causes a 0.68 unit increase in the tier1 capital ratio (equivalent to a change of 0.27 within-bank standard deviations) and a 11.9 unit decline in loan growth (equivalent to a decrease of 1.08 within-bank standard deviations). These results show that banks become more conservative by increasing capitalization and reducing lending in response to higher ratings.

A one-point increase in the rating also leads to a 0.46 point decrease in the bank's rating during the following exam. The magnitude of the effect is economically meaningful, equivalent to approximately 0.96 standard deviations in the within-bank variation in ratings. As noted previously, our finding that an exogenous higher rating causes a reduction in future ratings contrasts with the non-causal impact of ratings: An *endogenously* higher rating today can predict a higher rating for the bank next year because a higher rating reflects the examiner's prediction that the bank may face trouble in the future. An *exogenously* higher rating, *holding current bank fundamentals constant*, as our IV analysis establishes, causes a lower rating next year because the bank responds to the higher rating by taking conservative actions.

In Panel B, we examine the effect of bank ratings on loan growth and capital ratios two and three years into the future. Since the exogenous variation in the current CAMELS rating should be uncorrelated with future examiner assignment given the regular rotation structure, examiner discretion in the current year could have long-lived effects. We find that CAMELS ratings indeed have significant persistent effects (of approximately equal size) over the next two to three years, although the coefficients are more noisily estimated.

In Panel C, we explore the effect of CAMELS ratings on auxiliary outcomes, some of which function as quasi-placebo tests. We find that instrumented ratings have economically small and insignificant effects on the bank's non-performing loans (NPL) ratio and delinquency ratio. This is to be expected, because these ratios primarily depend on the performance of existing loans made *prior* to the determination of the current CAMELS rating. Exogenous changes in ratings that are not correlated with bank fundamentals should not strongly affect one-year ahead performance of loans, most of which were made prior to the ratings decision.

Altogether, these IV estimates show that changes in CAMELS ratings due to examiner discretion can have a large impact on bank capitalization and lending. If we assume that each bank

examination has an optimal outcome in terms of capitalization and lending, then examiner discretion will move these outcomes away from this optimum, creating potentially costly noise. These causal estimates also imply that examiner rotation, or any quasi-random assignment system of examiners to banks, can result in substantial volatility and uncertainty in bank outcomes, as bank activity will vary depending on which examiner is assigned for each examination and how that examiner chooses to exercise her personal discretion.

IV.E.1 Bank anticipatory response

So far, we have shown that discretion introduces variations in ratings that impact *ex post* bank capital ratios and lending behavior. In addition, discretion introduces uncertainty, which can potentially influence *ex ante* bank behavior. As modeled in Repullo (2024), bank managers who anticipate uncertainty and volatility in ratings may pre-emptively engage in conservative actions to reduce the probability of receiving an unsatisfactory CAMELS rating in the future.

To test for the *ex ante* influence of ratings driven by discretion, we exploit variation in discretionary uncertainty across states and over time. We measure discretionary uncertainty in a given state-year as the average level of absolute discretion and the standard deviation of directional discretion over the past five years. Higher values of these measures indicate greater uncertainty about future bank ratings. We then test whether banks in state-years with recent high discretionary activity respond by engaging in precautionary measures. The intuition behind this test is as follows: although bank management cannot directly observe the CAMELS ratings of other banks in their state, they can observe publicly disclosed actions—such as changes in capitalization and lending. If management sees substantial shifts in these behaviors across banks over time, they may infer that regulatory assessments in their state are uncertain and volatile.

We measure state-level absolute discretion as the average exam-level absolute discretion for all exams in the state in the past five years, excluding the current bank. We measure the state-level standard deviation in directional discretion as the standard deviation of exam-level direction discretion for all exams in the state in the past five years, excluding the current bank. Note that our measure of state-time-level absolute discretion captures uncertainty that is, by definition, uncorrelated with how tough examiners are on average within a state. The reason is that state-level directional discretion is a residual and *has an average of zero within a state quarter, by construction.* We caution that these tests of banks' anticipatory responses are intended to be suggestive rather than conclusive. Unlike our earlier IV analysis of *ex post* distortions in bank behavior, we lack exogenous variation in which states are subject to greater discretionary uncertainty. To better isolate the effects of uncertainty, we control for the bank's own most recent rating, observable bank characteristics, state fixed effects, and year-quarter fixed effects.

In Table 11 Panel A, we show that higher average state absolute discretion and state standard deviation of directional discretion is associated with significantly higher tier 1 capital ratios and lower loan growth. In terms of magnitudes, the pre-emptive response for loan growth is larger than the response for capital ratios (an interquartile range in state-level average examiner absolute discretion

is associated with banks reducing loan growth by 0.37 units and increasing capital ratios by 0.06 units, a 6% decline and a 1% increase relative to the sample medians, respectively).

In auxiliary regressions (Panel B), we find no effect of state-level discretion on NPL or delinquency ratios. This is expected because banks are unlikely to have a large degree of control over the performance of existing loans. Together, Table 11 suggests that discretion can impact both the *ex post* and *ex ante* behavior of banks.

IV.F Does discretion aid in making predictions?

While allowing for examiner discretion can inject randomness and uncertainty into bank regulation (because ratings vary simply due to examiner assignment), discretion also allows examiners to process soft information and to draw upon their experience and intuition when deciding upon ratings. As a result, it may also lead to more accurate assessments of bank quality and more accurate predictions of future bank outcomes, a key output of bank supervision.

We face two challenges in assessing the predictive power of discretionary ratings. First, measuring the "quality" of ratings is hampered by the fact that we do not know exactly what regulators seek to predict (and over what horizon). Indeed, regulators themselves may disagree over the exact objectives of the rating system. To evaluate rating quality, we assess the extent to which CAMELS ratings predict future ratings and future changes in bank observables that are strong predictors of CAMELS scores. While this approach may not capture all regulatory goals, it offers a reasonable proxy for a rating's ability to forecast future bank health.

Second, we are interested in the ability of a CAMELS rating to predict future bank outcomes. However, as we describe in detail in the framework in Section III.E, CAMELS ratings causally affect future bank outcomes in addition to predicting them. For example, a high CAMELS rating may predict that a bank's risk will increase but may also cause reduction in bank risk because the high CAMELS rating causes the bank to engage in more conservative lending. In our previous analysis, we isolated the causal effect of ratings by employing a leave-out-mean IV analysis. Now, to isolate the predictive power of ratings, we test how ratings predict near term changes in future bank outcomes that are unlikely to be affected by the ratings: the change in the next quarter of the performance of loans made *prior* to the exam.

We also assess the extent to which CAMELS ratings can predict the bank's future rating, usually assessed one year in the future. Here, the predictive and causal channels of the rating clearly go in opposite directions. A higher CAMELS rating may predict a high future CAMELS rating but should cause the bank to engage in risk-reduction, leading to a lower rating. Indeed, we estimated a negative causal impact of the current rating on the future rating, as shown in Table 10.

We measure the discretionary component of each exam rating as the exam-level $Directional\_Discretion_{ijrt}$ (defined as the actual rating minus the predicted rating). In Table 12 Panel A, we first test whether discretion is effective in predicting future bank outcomes. The dependent variables are changes in bank outcomes in the quarter after the rating relative to the same outcome in the quarter before the rating is released. We find that a higher value for the discretionary

component of each exam rating predicts significantly increased NPL ratio and delinquency ratio. Further, the discretionary component of ratings also predicts higher ratings in the next exam. Overall, we find that the discretionary component of ratings predicts future ratings and changes in bank outcomes that all head in the direction of the bank becoming less safe. These results suggest that examiners on average effectively use discretion to predict changes in bank outcomes that correspond with the bank becoming less safe in the near future.

However, finding that discretion helps to predict bank risk *on average* does not imply that more discretion is always better in making such predictions. In Panel B, we explore how the predictive power of discretionary component of ratings varies with examiner-level measures of absolute discretion. As before, we measure the discretionary component of each rating as the actual composite rating minus the predicted value based on bank observable hard information. The dependent variables are again changes in bank outcomes in the next quarter or the bank's next rating. We regress these bank outcomes on exam-level discretion interacted with three indicators for the level of examiner absolute discretion (calculated excluding the current observation). There is no omitted category, so each coefficient measures the extent to which the discretionary component of the rating is able to predict future changes in bank outcomes among the subset of examiners with low, medium, or high absolute discretion. We continue to find that the discretionary component of ratings for each exam predicts changes in bank outcomes that head in the direction of the bank becoming less safe. However, examiners in the top tercile of discretion do not produce more predictive ratings than those in the middle tercile.[27] In other words, greater discretion adds noise and uncertainty without improving forecast accuracy.

Finally, we compare the ability of the composite rating chosen by examiners and three counterfactual ratings to predict near-term future bank outcomes. The first counterfactual rating is the *predicted rating*, equal to the predicted value from a regression of actual ratings on bank observable hard information. The second counterfactual rating is the *reweighted composite rating*, where the weights for each component rating are chosen as the set of weights that provide the best estimate of the bank's actual composite rating in the next year (estimated from a regression of the bank's future rating on current component ratings, subject to the constraint that the coefficients sum to one). The counterfactual weights for components C, A, M, E, L, and S are 0.114, 0.185, 0.298, 0.114, 0.150, and 0.139, respectively. The third counterfactual rating is the composite rating if all components are equally weighted. To allow for comparison between the actual and counterfactual ratings, we round the counterfactual ratings to the nearest integer between 1 and 5.

These counterfactual ratings were chosen to test ideas from the literature on algorithm aversion (e.g., Dawes, 1979; Dawes, Faust, and Meehl, 1989), which has argued that humans attach too much weight to their personal insights and exercise too much discretion. Past research on algorithm aversion has argued that even extremely simple linear combinations of observable variables with unit coefficients can outperform human forecasts.

[27]The p-values for tests of equality shown at the bottom of the table also imply that examiners in the top tercile of discretion do not significantly outperform examiners in the bottom tercile for three out of the four outcomes.

As discussed in the Introduction, we purposely do not "race" examiner ratings against an advanced and complex algorithmic model (for an example of such a comparison, see e.g., Kleinberg et al. (2018a)). Given the extremely rapid rate of progress in AI and machine learning methods, such a test is unlikely to be informative in our setting. Even if human decisions *currently* outperform the best available algorithm, it is not obvious that humans would outperform the best available algorithm in the near future. Nevertheless, reliance on human decision making is likely to persist in banking regulation and many other institutions in the near future. Thus, we instead test whether imposing modest restrictions on human discretion can improve predictive power, while reducing noise.

Our first counterfactual removes human judgment of soft information and only uses observable hard information. The reweighted and equal-weight counterfactuals remove examiner discretion over how to weight each component rating. In particular, these reweighted composite ratings reduce the weight examiners place on subjective assessments of management quality from approximately 50% to 29% and 16.7%, respectively. Panel A of Table 13 shows the correlations between the real and three counterfactual ratings. While strongly positively correlated, there exists variation between the real and counterfactual ratings.

Panels B of Table 13 compare the ability of the actual and counterfactual ratings to predict proxies for future bank health. Panel B shows the area under the curve (AUC) of various receiver operating characteristic (ROC) curves. Outcomes are measured as an indicator for whether the bank's rating in the following year is unsatisfactory ($\geq$3), whether the change in the non-performing loan ratio in the quarter after the exam relative to the quarter before the exam is the top 10% of the sample, and whether the change in the loan delinquency ratio in the quarter after the exam relative to the quarter before the exam is in the top 10% of the sample. The 10% cutoff is chosen to approximately match the percentage of banks rated as unsatisfactory in the overall sample. Higher AUC values correspond to greater predictive power. Note that these loan portfolio outcomes are measured in changes relative to their levels just before the current exam and are thus difficult to forecast, as evidenced by the low AUCs across all four types of ratings.

We find that the actual rating is a better predictor of changes in bank quality than the predicted rating based on observable bank characteristics. However, the two reweighted counterfactual ratings outperform the actual rating. Both reweighted composite ratings are stronger predictors of the bank's next rating and of near-term changes in the performance of the bank's existing loan portfolio. The performance gain of the two reweighted composite ratings over the actual rating amounts to only a small increase in AUC. Nevertheless, like our earlier results in Table 12, these results suggest placing moderate limitations on examiner discretion can lead to predictions that are at least as powerful, while simultaneously reducing noise.

Overall, our findings suggest that allowing for examiner discretion can lead to more predictive and accurate forecasts of bank health. Our analysis does not support an extreme view of algorithm aversion in which simple linear combinations of observable hard information outperforms human-generated predictions. However, our results also show that forecasts could potentially be improved

by limiting the degree of discretion that examiners hold over how component ratings are weighted. In particular, constraining the way in which component ratings are aggregated to form the overall composite rating may lead to weakly better forecasts of changes in loan performance while simultaneously reducing noise. Likewise, our earlier results in Table 12 showed that examiners who choose to exercise less discretion generate ratings that are just as predictive as those who exercise more discretion.

V Conclusion

Using detailed data on the supervisory decisions of US banking regulators, we find that professional bank examiners exercise significant personal discretion. Quasi-random assignment of examiners to banks guarantees fairness, in that no bank systematically faces more lenient regulation. However, human discretion injects a large degree of noise and uncertainty into the system.

We develop a framework and provide evidence showing that human discretion in decision-making arises from how individuals weight component issues. Final decisions can be modeled as weighted sums of ratings across these components, and we identify three key sources of disagreement. First, decision-makers place disproportionate weight on the most subjective components—examiners in our setting assign 50% of the weight to subjective assessments of management quality. Second, they vary in how they weight components, leading to disagreement in overall ratings even when they agree on individual parts. Third, they treat relatively objective factors as if they were highly subjective. We show that the resulting examiner disagreement has a large and persistent causal impact on future bank capitalization and credit supply, contributing to volatility and uncertainty in outcomes. Banks respond conservatively in anticipation. While replacing discretion with a simple algorithm reduces predictive power, placing moderate limits on discretion enhances the signal quality of supervisory ratings.

It is worth noting that the optimal amount of discretion is unlikely to be zero because discretion facilitates the use of soft information. Indeed, we find that the discretionary component of decisions does have predictive power. On the other hand, human discretion leads to arbitrary regulation and noise, and increases uncertainty and volatility. As pointed out by Kahneman, Sibony, and Sunstein (2021), a key advantage of algorithms, even if the algorithms fail to outperform human predictions, is that algorithms do not introduce as much uncertainty and volatility.

More broadly, some amount of uncertainty around bank supervisory models such as stress tests may be desirable in that it could limit opportunistic gaming by banks and encourage conservative actions, which, depending on one's beliefs, may be a desirable outcome (e.g., Leitner and Williams, 2022). It is important to note, however, that uncertainty added due to examiner discretion is very different from opacity around the stress test model. In stress tests, banks face uncertainty around the true model, but regulators know the true model and can fully control test outcomes as a function of inputs. In contrast, individual examiner discretion induces uncertainty for both the regulatory authority and banks, and the impact of discretion on outcomes may be harder to discern

for regulators. The optimal amount of human discretion and its governance for each type of system, including banking regulators (e.g., Laeven and Levine, 2009), is left for future research.

References

Agarwal, Sumit, David Lucca, Amit Seru, and Francesco Trebbi. 2014. "Inconsistent Regulators: Evidence from Banking". *Quarterly Journal of Economics*, 129(2), 889–938.

Anderson, Norman H. 1971. "Integration Theory and Attitude Change". *Psychological Review*, 78, 171–206.

Angelova, Victoria, Will Dobbie, and Crystal S. Yang. 2022. "Algorthmic Recommendations and Human Discretion". Working Paper.

Angrist, Joshua D. and Guido W. Imbens. 1994. "Identification and Estimation of Local Average Treatment Effects". *Econometrica*, 62(2), 467–475.

Arnold, David, Will Dobbie, and Peter Hull. 2022. "Measuring Racial Discrimination in Bail Decisions". *American Economic Review*, 112(9), 2992–3038.

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions". *Quarterly Journal of Economics*, 133(4), 1885–1932.

Benson, Alan, Danielle Li, and Kelly Shue. 2019. "Promotions and the peter principle". *The Quarterly Journal of Economics*, 134.4, 2085–2134.

Benson, Alan, Danielle Li, and Kelly Shue. 2021. ""Potential" and the gender promotion gap". Working Paper.

Berger, Allen N., Robert DeYoung, Hesna Genay, and Gregory F. Udell. 2000. "Globalization of financial institutions: Evidence from cross-border banking performance". *Brookings-Wharton papers on financial services 2000*, no. 1, 23–120.

Bloom, Nicholas. 2009. "The impact of uncertainty shocks". *Econometrica*, 77(3), 623–685.

Bohren, Aislinn, Kareem Haggag, Alex Imas, and Devin Pope. 2023. "Inaccurate statistical discrimination: An identification problem". Forthcoming, *Review of Economics and Statistics*.

Bris, Arturo, Ivo Welch, and Ning Zhu. 2006. "The Costs of Bankruptcy: Chapter 7 Liquidation versus Chapter 11 Reorganization". *The Journal of Finance*, 61(3), 1253–1303.

Calomiris, Charles and Gary Gorton. 1991. "The Origins of Banking Panics: Models, Facts, and Bank Regulation". NBER Publication on Financial Markets and Financial Crises, University of Chicago Press.

Calomiris, Charles W. 2006. "The Regulatory Record of the Greenspan Fed". *American Economic Review Papers and Proceedings*, 96, 170–173.

Chan, David C., Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists". *Quarterly Journal of Economics*, 137(2), 729–784.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson. 2016. "Health care exceptionalism? Performance and allocation in the US health care sector". *American Economic Review*, 106(8), 2110–2144.

Chang, Tom and Antoinette Schoar. 2006. "Judge Specific Differences in Chapter 11 and Firm Outcomes". Working Paper.

Chilton, Adam S. and Marin K. Levy. 2015. "Challenging the randomness of panel assignment in the Federal Courts of Appeals". *Cornell Law Review*, 101.

Dawes, Robyn M. 1979. "The Robust Beauty of Improper Linear Models in Decision Making". *American Psychologist*, 34(7), 571–582.

Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. "Clinical versus actuarial judgment". *Science*, 243(4899), 1668–1674.

Dobbie, William and Jae Song. 2015. "Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection". *American Economic Review*, 105(3), 1272–1311.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramodarai, and Ansgar Waither. 2022. "Predictably Unequal? The Effects of Machine Learning on Credit Markets". *Journal of Finance*, 77(1), 5–47.

Garicano, Luis. 2012. "Five lessons from the Spanish cajas debacle for a new euro-wide supervisor". *Banking Union for Europe*, 79.

Gissler, Stefan, Jeremy Oldfather, and Doriana Ruffino. 2016. "Lending on hold: Regulatory uncertainty and bank lending standards". *Journal of Monetary Economics*, 81, issue C, 89–101.

Hall, Matthew. 2010. "Randomness Reconsidered: Modeling Random Judicial Assignment in the US Courts of Appeals". *Journal of Empirical Legal Studies*, 7(3), 574–589.

Hirtle, Beverly J. and Jose A. Lopez. 1999. "Supervisory Information and the Frequency of Bank Examinations". *Economic Policy Review*, 5, 1–19.

Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li. 2018. "Discretion in Hiring". *Quarterly Journal of Economics*, 133(2), 765–800.

Huang, Laura and Jone L. Pearce. 2015. "Managing the Unknowable: The Effectiveness of Early-Stage Investor Gut Feel in Entrepreneurial Investment Decisions". *Administrative Science Quarterly*, 60(4), 634–670.

Huber, Kilian. 2021. "Are bigger banks better? Firm level evidence from Germany". *Journal of Political Economy*, 129(7), 2023–2066.

Iyer, Rajkamal, Asim I. Khwaja, Erzo F.P. Luttmer, and Kelly Shue. 2016. "Screening peers softly: Inferring the quality of small borrowers". *Management Science*, 62(6), 1554–1577.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*, Little, Brown Spark.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human decisions and machine predictions". *The Quarterly Journal of Economics*, 133(1), 237–293.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human decisions and machine predictions". *The Quarterly Journal of Economics*, 133, 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh R. Kleinberg. 2018. "Algorithmic fairness". *Aea papers and proceedings*, 108, 22–27.

Laeven, Luc and Ross Levine. 2009. "Bank Governance, Regualtion and Risk Taking". *Journal of Financial Economics*, 93(2), 259–275.

Leitner, Yaron and Basil Williams. 2022. "Model Secrecy and Stress Tests". *Journal of Finance*, forthcoming.

Levine, Timothy R., Steven A. McCornack, and Hee Sun Park. 1999. "Accuracy in detecting truths and lies: Documenting the "veracity effect"". *Communication Monographs*, 66(2), 125–144.

Liberti, Jose M. and Atif R. Mian. 2009. "Estimating the Effect of Hierarchies on Information Use". *Review of Financial Studies*, 22(10), 4057–4090.

Lipsky, Michael. 2010. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*, 30th Anniversary Expanded Edition.

Meehl, Paul E. 1954. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*, University of Minnesota Press.

Morris, Carl N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications". *Journal of the American Statistical Association*, 78(381), 47–55.

Mullainathan, Sendhil and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care". *Quarterly Journal of Economics*, 137(2), 679–727.

Petersen, Mitchell A. and Raghuram G. Rajan. 1994. "The benefits of lending relationships: Evidence from small business data". *The Journal of Finance*, 49(1), 3–37.

Petersen, Mitchell A. and Raghuram G. Rajan. 2002. "Does distance still matter? The information revolution in small business lending". *Journal of Finance*, 57(6), 2533–2570.

Rajan, Uday, Amit Seru, and Vikrant Vig. 2015. "The failure of models that predict failure: Distance, incentives, and defaults". *Journal of Financial Economics*, 115(2), 237–260.

Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Philip G. Schrag. 2010. "Refugee roulette: Disparities in asylum adjudication". *The Modern Law Review*, 73(4), 679–682.

Repullo, Rafael. 2024. "Regulation, Supervision, and Bank Risk-Taking". Working Paper.

Sampath, Bhaven and Heidi Williams. 2019. "How do Patents Affect Follow-On Innovation? Evidence from Human the Genome". *American Economic Review*, 109(1), 203–236.

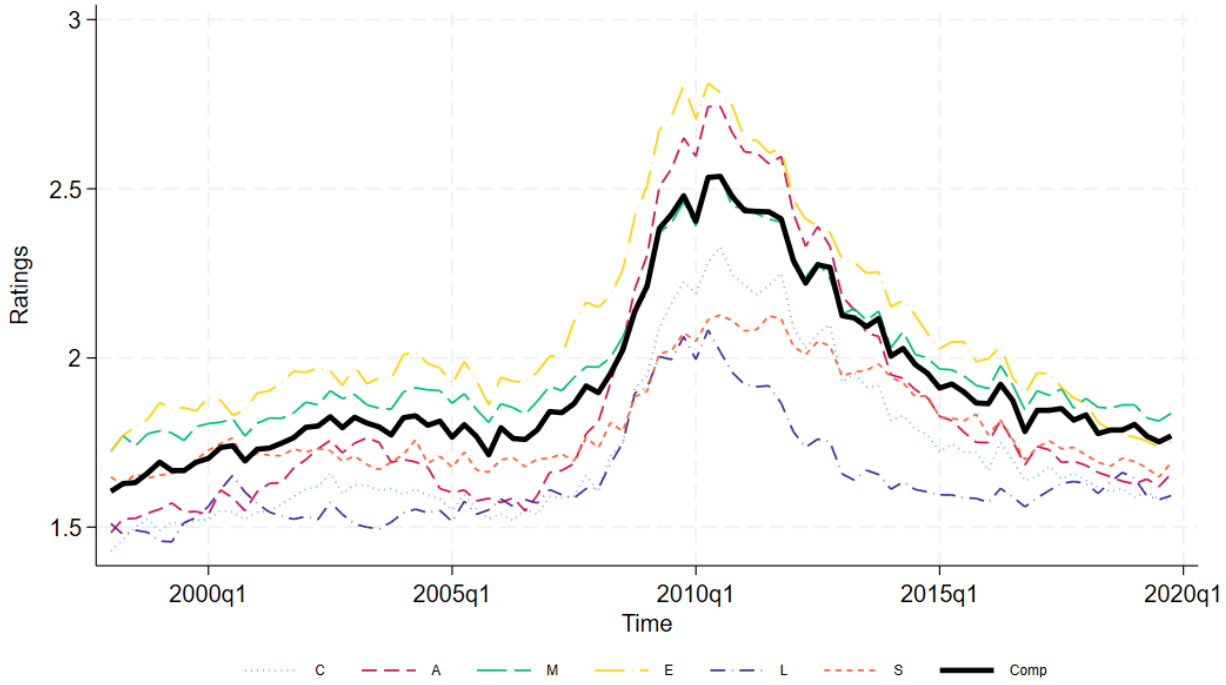
Shleifer, Andrei and Robert W. Vishny. 1999. "The Grabbing Hand: Government Pathologies and Their Cures". Cambridge, MA, Harvard University Press.

Stein, Jeremy C. 2002. "Information production and capital allocation: Decentralized versus hierarchical firms". *Journal of Finance*, 57(5), 1891–1921.

Tversky, Amos and Daniel Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases". *Science*, 185(4157), 1124–1131.

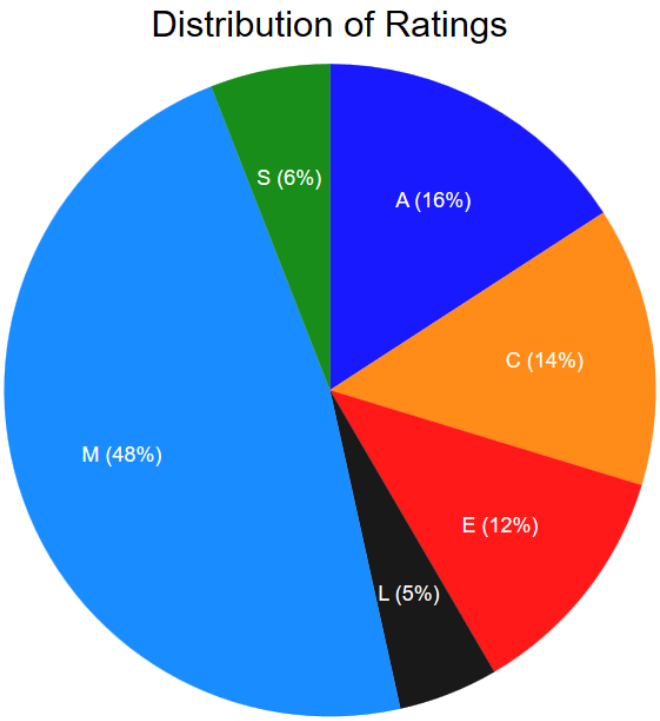
Yang, Crystal S. 2015. "Free at Last? Judicial Discretion and Racial Disparities in Federal Sentencing". *The Journal of Legal Studies*, 44(1), 75–111.

Figure 1: Evolution of ratings over time



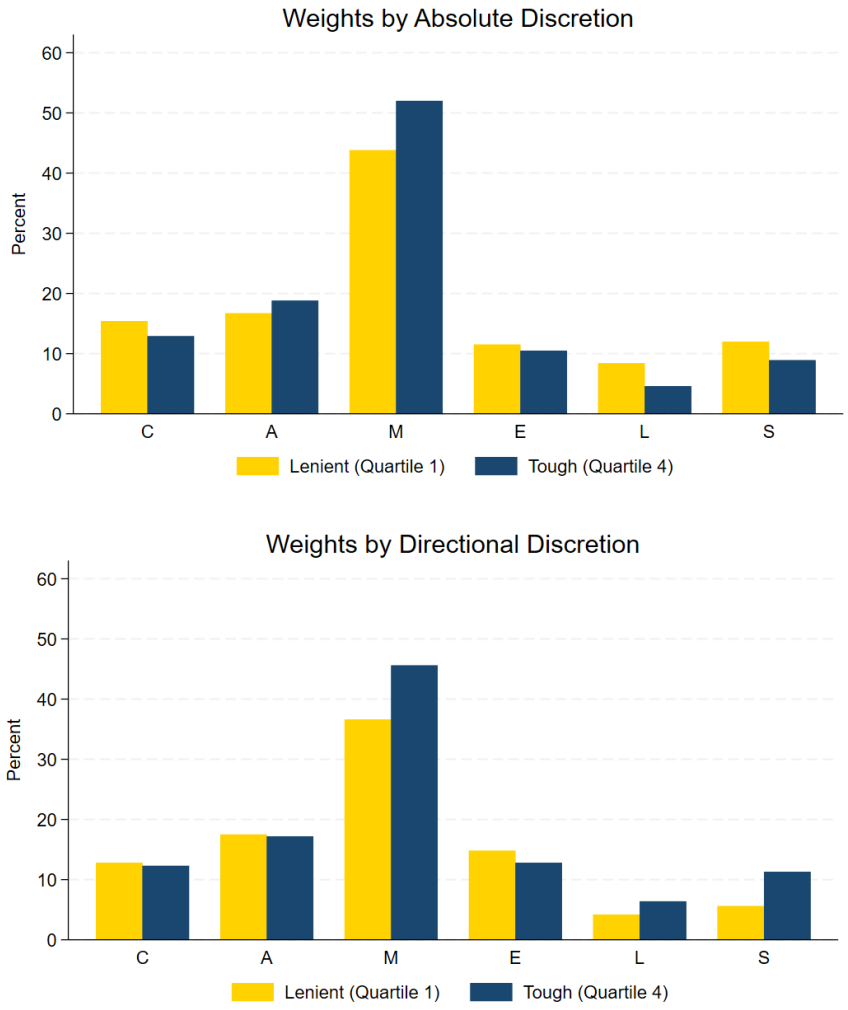
This figure shows the average composite and component CAMELS ratings across all exams in our data sample over time. The composite rating is a summary measure of component ratings: capital, assets, management, earnings, liquidity, and sensitivity to market risk, which together form the acronym CAMELS. Examiners have some degree of discretion over each component rating as well as over how component ratings are aggregated to form the composite rating. The composite and component ratings range from 1 to 5 with higher ratings representing greater safety and soundness concerns.

Figure 2: Weights on component ratings



This figure shows the composite rating as a weighted sum of the component ratings, using the estimates from the OLS regression of the composite rating against component ratings, with the constraint that the coefficients sum to one. Results are also reported in Table 7 Panel A Column 2.

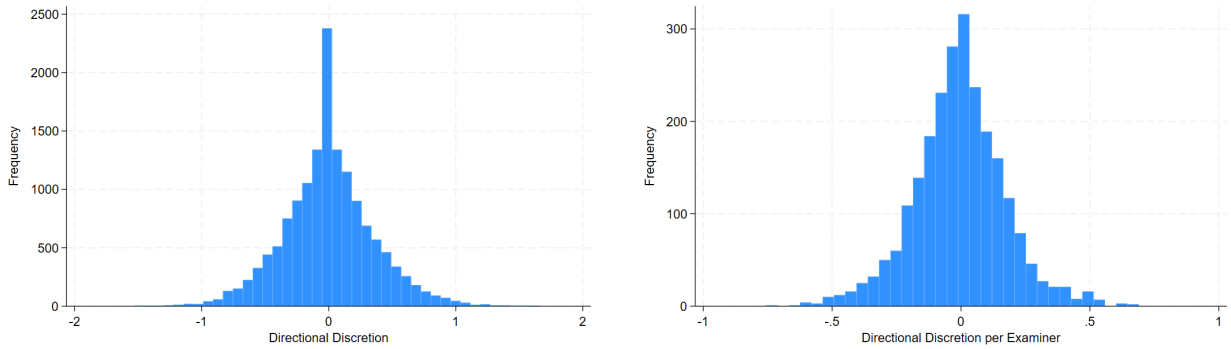
Figure 3: Heterogeneity in weights by examiner discretion



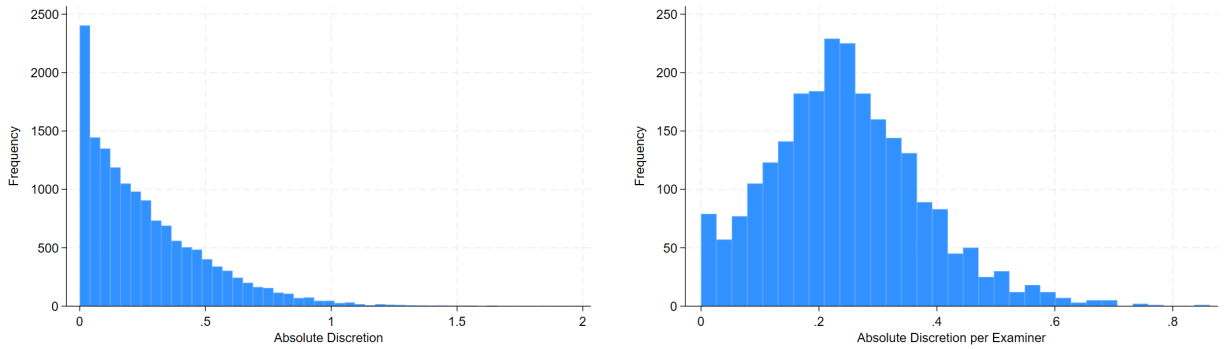
This figure shows the composite rating as a weighted sum of the component ratings, separately for observations in the top and bottom quartiles of absolute discretion (top panel) and top and bottom quartiles of directional discretion (bottom panel). Detailed corresponding regression estimates are reported in Table 9.

Figure 4: Distribution of directional discretion and absolute discretion

Panel A: Distribution of directional discretion at the exam and examiner levels.



Panel B: Distribution of absolute discretion at the exam and examiner levels.



These figures show the distribution of estimates of directional discretion (Panel A) and absolute discretion (Panel B) at the exam level (left side) and examiner level (right side). Directional discretion is the average signed value of discretion, defined as the residual component of the rating that cannot be explained by observable bank characteristics. It captures whether an examiner tends to be tougher or more lenient than predicted by observable hard information. Absolute discretion, by contrast, is the average of the absolute values of discretion. It reflects how much an examiner relies on case-specific soft information, gut feelings, or personal judgment, regardless of direction.

Table 1: Summary statistics

This table presents summary statistics of our data. Panel A describes examiner experience in terms of number of exams conducted within our sample or the number of years present within our sample. Panel B describes the distribution of each component of the CAMELS rating, as well as the composite rating. Panel C shows the transition probabilities between the current and next composite ratings for banks over time. Panel D describes bank observables as of the end of the quarter immediately before the examiner's rating is finalized. *Tier1 ratio* is defined as tier1 capital / risk-adjusted assets. *Loan growth* is defined as log(loans) – log(loans one year ago). *Leverage* is defined as tier1 capital / assets. *ROA* is defined as net income / assets. *NPL ratio* is defined as loans >90 days past due + non-accrual loans / total loans. *Delinq ratio* is defined as delinquent loans / loans. *Efficiency* is defined as non-interest expenses / revenue. All ratios are computed as percents, with numerators and denominators measured in thousands of $US. *Bank assets* are reported in millions $US.

Panel A: Examiner experience

	N	Mean	S.D.	Min	p25	p50	p75	Max
Examiner-level								
Number of exams	2,407	6.1	5.7	2	3	4	7	74
Examiner-exam-level								
Years experience	14,679	5.4	4.7	1	2	4	7	23
Number of exams so far	14,679	6.2	6.8	1	2	4	8	74

Panel B: CAMELS rating components

Rating	N	Mean	S.D.	S.D.(Within bank)	Min	p25	p50	p75	Max	Frac ≥ 3
Capital	14,679	1.74	0.76	0.45	1	1	2	2	5	0.09
Assets	14,679	1.81	0.88	0.60	1	1	2	2	5	0.16
Management	14,679	1.96	0.77	0.49	1	1	2	2	5	0.16
Earnings	14,679	2.11	1.01	0.62	1	1	2	2	5	0.27
Liquidity	14,679	1.68	0.69	0.43	1	1	2	2	5	0.08
Sensitivity	14,679	1.82	0.64	0.41	1	1	2	2	5	0.09
Composite	14,679	1.93	0.77	0.48	1	1	2	2	5	0.15

Panel C: Transitions probability matrix

	Next Rating					
Rating	1	2	3	4	5	Total
1	79.8	19.9	1.0	0.1	0.0	100
2	8.6	83.6	6.5	1.1	0.2	100
3	0.0	36.4	53.8	8.4	1.4	100
4	0.0	2.5	18.7	58.1	20.7	100
5	0.0	0.0	3.3	10.0	86.7	100
Total	27.5	59.9	9.5	2.3	0.8	100

Panel D: Bank observables

	N	Mean	S.D.	S.D.(within bank)	Min	p25	p50	p75	Max
Tier1 ratio	14,679	9.72	2.55	2.49	6.08	7.96	9.17	10.77	17.70
Loan growth	14,679	9.13	15.30	10.99	-16.39	0.29	6.53	14.24	63.14
Leverage	14,679	10.35	2.73	1.31	6.34	8.46	9.80	11.59	18.57
ROA	14,679	0.28	0.28	0.14	-0.75	0.18	0.32	0.44	0.73
NPL ratio	14,679	1.31	1.61	1.03	0.00	0.27	0.75	1.66	7.28
Delinq ratio	14,679	2.20	2.25	1.48	0.00	0.64	1.47	2.95	9.89
Efficiency	14,679	56.91	20.37	16.08	27.19	42.38	53.80	66.67	124.96
Assets(millions)	14,679	2,267	15,912	3696.8	3	82	186	500	621,240

Table 2: Measuring discretion and assessing the predictive power of examiner leave-out-mean discretion

Column 1 presents a regression of the composite rating on bank characteristics at the time of the exam. Column 2 presents a regression of the composite rating for the current exam on the examiner's leave-out-mean directional discretion, a jack-knife instrument measuring each examiner's bias, as well as bank characteristics at the time of the exam. Standard errors are in parentheses and clustered by examiner. $*$, $**$, and $***$ denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Composite rating	(1)	(2)
Examiner directional discretion (LO)		0.170***
		(0.034)
Tier1 ratio	-0.023***	-0.049***
	(0.006)	(0.006)
Loan growth	-0.002***	-0.003***
	(0.000)	(0.000)
ROA	-0.512***	-0.669***
	(0.046)	(0.062)
NPL ratio	0.045***	0.072***
	(0.007)	(0.009)
Delinq Ratio	0.014***	0.051***
	(0.005)	(0.007)
Efficiency	-0.001	0.001
	(0.001)	(0.001)
Leverage	-0.017***	-0.006
	(0.005)	(0.006)
Observations	14,679	14,679
R-squared	0.797	0.798
Bank FE	Yes	Yes
Location-quarter FE	Yes	Yes

Table 3: Identifying assumption—Examiner rotation and bank health

This table tests whether bank characteristics at the time of the exam are correlated with examiner directional discretion and absolute discretion. It also tests whether changes in the performance of existing loans in the quarter after the exam relative to the quarter before the exam are correlated with examiner discretion. Panel A regresses bank characteristics on the examiner's leave-out-mean directional discretion, defined as the examiner's average directional discretion across other exams, excluding the current observation. Panel B regresses bank characteristics on the examiner's leave-out-mean absolute discretion, defined as the examiner's average absolute discretion across other exams, excluding the current observation. In Panel C, observations are ordered by bank-time, and we regress the leave-out-mean directional discretion and absolute discretion of the current examiner on the leave-out-mean directional discretion and absolute discretion of the examiner who previously examined the bank. Standard errors are in parentheses and clustered by examiner. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Rotation and examiner leave-out-mean direction discretion

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Tier1 ratio	Loan growth	Leverage	NPL ratio	Delinq ratio	Chg NPL Ratio	Chg Delinq ratio
Examiner dir disc (LO)	-0.230	-0.356	-0.242	0.192	0.156	-0.032	-0.066
	(0.149)	(0.763)	(0.161)	(0.143)	(0.123)	(0.045)	(0.067)
Observations	14,679	14,679	14,679	14,679	14,679	14,555	14,555
R-squared	0.264	0.306	0.258	0.398	0.425	0.260	0.279
Location-qtr FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Panel B: Rotation and examiner leave-out-mean absolute discretion

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Tier1 ratio	Loan growth	Leverage	NPL ratio	Delinq ratio	Chg NPL Ratio	Chg Delinq ratio
Examiner abs disc (LO)	0.008	-1.727	0.025	0.162	0.135	-0.017	0.015
	(0.227)	(1.264)	(0.244)	(0.133)	(0.187)	(0.072)	(0.104)
Observations	14,679	14,679	14,679	14,679	14,679	14,555	14,555
R-squared	0.263	0.306	0.358	0.418	0.425	0.260	0.279
Location-qtr FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Panel C: Autocorrelation in examiner assignment to banks

	(1)	(2)
	Examiner dir disc (LO)	Examiner abs disc (LO)
Prev. examiner dir disc (LO) (same bank)	-0.022	
	(0.015)	
Prev. Examiner abs disc (LO) (same bank)		-0.015
		(0.015)
Observations	8,742	8,742
R-squared	0.200	0.307
Location-quarter FE	Yes	Yes

Table 4: Directional discretion and absolute discretion

This table presents summary statistics of our estimates of directional discretion and absolute discretion, as defined in Section 3.2. We regress the composite CAMELS rating on the set of bank variables described in Panel C of Table 1. Using this regression, we create predicted ratings for each exam. The *non-integer* label implies that the predicted rating from the regression is allowed to be a continuous real number, while the *integer* label implies that the predicted rating from the regression is rounded to the nearest integer value, 1 through 5. For each exam-level observation, directional discretion is defined as the actual composite rating minus the predicted rating, and absolute discretion is defined as the absolute value of the directional discretion. For each examiner-level observation, we take the average of the examiner's directional discretion or absolute discretion across all exams associated with the examiner.

Panel A: Exam-level

Discretion Type	Pred rating	Obs	Mean	S.D.	p10	p25	p50	p75	p90
Directional	Non-integer	14,679	0.00	0.35	-0.43	-0.20	0.00	0.19	0.44
	Integer	14,679	0.00	0.39	0.00	0.00	0.00	0.00	0.00
Absolute	Non-integer	14,679	0.26	0.24	0.01	0.08	0.20	0.38	0.59
	Integer	14,679	0.15	0.36	0.00	0.00	0.00	0.00	1.00

Panel B: Examiner-level

Discretion Type	Pred rating	Obs	Mean	S.D.	p10	p25	p50	p75	p90
Directional	Non-integer	2,407	-0.01	0.18	-0.22	-0.11	0.00	0.10	0.20
	Integer	2,407	0.00	0.19	-0.25	0.00	0.00	0.00	0.21
Absolute	Non-integer	2,407	0.25	0.13	0.08	0.16	0.24	0.32	0.41
	Integer	2,407	0.14	0.18	0.00	0.00	0.00	0.25	0.40

Table 5: Disagreement in concurrent exams

This table examiners disagreement among examiners during concurrent bank examinations. We examine concurrent exams in which multiple examiners from different agencies assess the same bank in the same quarter. Panel A reports the fraction of concurrent exams in which the relevant CAMELS composite or component ratings differed. Panel B assesses the comparability between banks involved in concurrent exams and those in our baseline sample, presenting regression results of bank observable measures on an indicator for whether the exam was held concurrently with another exam conducted by another lead examiner. Standard errors in parentheses are clustered by bank. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Exam-level

Rating	Obs	Disagreement
Composite	410	0.28
Capital	410	0.19
Assets	410	0.23
Management	410	0.31
Earnings	410	0.19
Liquidity	410	0.13
Sensitivity	410	0.18

Panel B: Concurrent versus non-concurrent exams

	Tier1 ratio	Loan growth	Leverage	ROA	Efficiency	Delinq Ratio	NPL ratio
Concurrent	-0.213	0.915	0.213	-0.021	-0.449	0.274	0.312***
	(0.169)	(1.241)	(0.191)	(0.017)	(0.856)	(0.178)	(0.119)
Observations	14,679	14,679	14,679	14,679	14,679	14,679	14,679
R-squared	0.264	0.266	0.224	0.294	0.826	0.323	0.338
Agency FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 6: Predictability of examiner leave-out-mean discretion for component ratings

This table presents regression results examining how examiners' biases affect each component issue. Examiner directional discretion (LO) is the average discretion by examiner i for each relevant component (calculated excluding the current exam). The regression relates the current component rating to the examiner's average discretion for that component (calculated excluding the current exam), controlling for bank observables and state-year-quarter fixed effects. Standard errors are in parentheses and clustered by examiner. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

	C	A	M	E	L	S
Examiner directional discretion (LO)	0.126***	0.134***	0.190***	0.142***	0.096***	0.180***
	(0.035)	(0.036)	(0.034)	(0.037)	(0.037)	(0.036)
Tier1 ratio	-0.105***	-0.042***	-0.040***	-0.041***	-0.032***	-0.024***
	(0.006)	(0.008)	(0.007)	(0.008)	(0.007)	(0.006)
Loan growth	-0.003***	-0.005***	-0.003***	-0.002***	-0.001*	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Leverage	-0.024***	-0.005	-0.001	0.002	-0.029***	-0.017***
	(0.006)	(0.007)	(0.006)	(0.007)	(0.006)	(0.006)
ROA	-0.372***	-0.494***	-0.598***	-1.780***	-0.326***	-0.509***
	(0.064)	(0.066)	(0.066)	(0.075)	(0.067)	(0.060)
NPL ratio	-0.071***	-0.059***	-0.054***	-0.112***	-0.085***	-0.046***
	(0.018)	(0.020)	(0.019)	(0.021)	(0.018)	(0.017)
Delinq Ratio	0.082***	0.145***	0.095***	0.095***	0.073***	0.045***
	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)
Efficiency	0.004***	0.002**	0.001	-0.003***	0.000	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	14,679	14,679	14,679	14,679	14,679	14,679
R-squared	0.798	0.788	0.744	0.827	0.725	0.690
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes
Location-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes

Table 7: Weights on CAMELS rating components

This table shows examiners' weight on each component rating. Panel A shows how the composite rating is related to the component ratings. Column 1 presents an unconstrained OLS regression. Column 2 estimates the same regression subject to the constraint that the coefficients sum to one. Column 3 reports estimates from an ordered logit model. Panel B shows constrained regressions similar to that in Column 2 of Panel A, over the pre-recession, recession, and post-recession time periods. Standard errors in parentheses are clustered by examiner. $*$, $**$, and $***$ denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Weights			
	(1)	(2)	(3)
Composite rating	OLS	OLS Constrained	Ordered Logit
C-rat	0.147***	0.138***	3.099***
	(0.006)	(0.005)	(0.111)
A-rat	0.159***	0.157***	3.459***
	(0.005)	(0.004)	(0.100)
M-rat	0.482***	0.477***	5.280***
	(0.008)	(0.005)	(0.117)
E-rat	0.113***	0.122***	2.381***
	(0.004)	(0.003)	(0.078)
L-rat	0.069***	0.047***	1.677***
	(0.005)	(0.004)	(0.090)
S-rat	0.097***	0.059***	1.971***
	(0.006)	(0.004)	(0.091)
Observations	14,679	14,679	14679
R-squared	0.889		

Panel B: Weights over time

	(1)	(2)	(3)
Composite rating	1998Q1-2007Q4	2008Q1-2012Q4	2013Q1-2020Q1
C-rat	0.117***	0.162***	0.147***
	(0.007)	(0.010)	(0.009)
A-rat	0.137***	0.212***	0.117***
	(0.006)	(0.009)	(0.008)
M-rat	0.470***	0.433***	0.544***
	(0.007)	(0.011)	(0.009)
E-rat	0.136***	0.117***	0.080***
	(0.005)	(0.007)	(0.006)
L-rat	0.069***	0.017*	0.040***
	(0.006)	(0.009)	(0.007)
S-rat	0.071***	0.059***	0.072***
	(0.006)	(0.009)	(0.008)
Constrained regression	Yes	Yes	Yes
Observations	7,291	3,345	4,043

Table 8: Heterogeneity in weights across examiners

This table displays the heterogeneity in component weights across examiners. We restrict the sample to 415 lead examiners who are associated with at least 10 exams. For each examiner, we regress the composite rating on the component ratings to recover examiner-specific component weights. The table reports the average and standard deviation of the weights assigned to each component across examiners.

Component	Obs	Average weight	Std dev. Weights
Capital	415	0.152	0.161
Assets	415	0.158	0.141
Management	415	0.478	0.174
Earnings	415	0.106	0.116
Liquidity	415	0.078	0.151
Sensitivity	415	0.081	0.163

Table 9: Heterogeneity in weights by examiner discretion

This table shows how weights for component ratings vary with the examiner's directional discretion and absolute discretion. Column 1 restricts the sample to observations corresponding to examiners in the top and bottom quartile of directional discretion, with quartile 4 representing examiners that choose tougher (i.e., higher) ratings on average. We regress the composite rating on the component ratings and the interaction between the component ratings and an indicator for quartile 4. The coefficients on the component ratings represent the weights for each component for examiners in quartile 1. The coefficients on the interaction terms represent the difference in weights for quartile 4 relative to quartile 1. Column 2 estimates the same regression, with quartiles sorted by examiner absolute discretion, with quartile 4 representing examiners who exercise more absolute discretion on average. Standard errors in parentheses are clustered by examiner. $*$, $**$, and $***$ denote statistical significance at the 10%, 5%, and 1% levels, respectively.

	(1)		(2)
Composite rating	Directional discretion		Absolute discretion
C-rat	0.128***	C-rat	0.154***
	(0.011)		(0.011)
C-rat × $Q4_{\text{high direct disc}}$	-0.005	C-rat × $Q4_{\text{high abs disc}}$	-0.025*
	(0.013)		(0.014)
A-rat	0.175***	A-rat	0.167***
	(0.009)		(0.010)
A-rat × $Q4_{\text{high direct disc}}$	-0.003	A-rat × $Q4_{\text{high abs disc}}$	0.021
	(0.013)		(0.013)
M-rat	0.366***	M-rat	0.438***
	(0.014)		(0.016)
M-rat × $Q4_{\text{high direct disc}}$	0.090***	M-rat × $Q4_{\text{high abs disc}}$	0.082***
	(0.019)		(0.021)
E-rat	0.148***	E-rat	0.115***
	(0.006)		(0.007)
E-rat × $Q4_{\text{high direct disc}}$	-0.020**	E-rat × $Q4_{\text{high abs disc}}$	-0.010
	(0.010)		(0.010)
L-rat	0.042***	L-rat	0.084***
	(0.009)		(0.009)
L-rat× $Q4_{\text{high direct disc}}$	0.022*	L-rat × $Q4_{\text{high abs disc}}$	-0.038***
	(0.012)		(0.013)
S-rat	0.056***	S-rat	0.120***
	(0.009)		(0.011)
S-rat × $Q4_{\text{high direct disc}}$	0.057***	S-rat × $Q4_{\text{high abs disc}}$	-0.031**
	(0.012)		(0.013)
Observations	7,339		7,339
R-squared	0.922		0.904

Table 10: Causal impact of discretion in ratings

This table presents estimates of the causal effect of exogenously higher composite ratings due to examiner discretion on *ex post* bank outcomes. The results derive from 2SLS estimates in which the composite rating is instrumented with the examiner's leave-out-mean directional discretion (the first stage regression is as reported in Column 3 of Table 2). In Panel A, bank outcomes are measured four quarters after the quarter in which the current exam rating is finalized, or as of the next exam in Column 1 and 2. Panel B examines the impact on bank outcomes two and three years after the current exam. Panels A and B examine outcomes likely to be directly affected by exogenous changes in CAMELS ratings, while Panel C presents other auxiliary bank outcomes. Bank controls consist of variables reported in Table 1 Panel D. Standard errors in parentheses are adjusted for the two-step instrumental variables procedure and clustered by examiner. $*$, $**$, and $***$ denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Full sample				
	(1)	(2)	(3)	(4)
Future bank outcome	Next rating	Troubled Bank	Tier1 ratio - 1yr	Loan growth - 1yr
Pred composite rating	-0.458*	-0.121**	0.680*	-11.922***
	(0.274)	(0.054)	(0.395)	(4.172)
Observations	12,802	12,802	13,040	13,033
Bank controls	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes
Location-quarter FE	Yes	Yes	Yes	Yes
Panel B: Long run impact				
	(1)	(2)	(3)	(4)
Future bank outcome	Tier1 ratio - 2yr	Loan growth - 2yr	Tier1 ratio - 3yr	Loan growth - 3yr
Pred composite rating	1.041**	-11.993*	0.861*	-21.660***
	(0.498)	(6.627)	(0.489)	(8.381)
Observations	12,298	12,304	11,709	11,712
Bank controls	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes
Location-quarter FE	Yes	Yes	Yes	Yes
Panel C: Auxiliary outcomes				
		(1)		(2)
Future bank outcome		NPL ratio		Delinq ratio
Pred composite rating		0.136		0.241
		(0.120)		(0.551)
Observations		13,032		13,032
Bank controls		Yes		Yes
Bank FE		Yes		Yes
Location-quarter FE		Yes		Yes

Table 11: Anticipatory bank response

This table regresses bank outcomes on state-level discretionary uncertainty, as proxied by the average level of absolute discretion and the standard deviation of directional discretion over the past five years. Panel A presents bank outcomes that are likely to be targeted and important determinants of the CAMELS rating. Panel B presents other auxiliary bank outcomes. Bank controls consist of variables reported in Table 1 Panel D. Standard errors in parentheses are clustered by bank. ∗, ∗∗, and ∗∗∗ denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Pre-emptive capital ratio and loan growth response				
	(1)	(2)	(3)	(4)
	Tier1 ratio	Tier1 ratio	Loan growth	Loan growth
State absolute discretion	0.82*		-4.76**	
	(0.43)		(1.88)	
State sd direction discretion		0.63*		-3.20**
		(0.35)		(1.54)
Observations	14,237	14,228	14,067	14,067
R-squared	0.47	0.46	0.12	0.11
Quarter FE	Yes	Yes	Yes	Yes
Lagged Bank Controls	Yes	Yes	Yes	Yes
Panel B: Auxiliary Outcomes				
	(1)	(2)	(3)	(4)
	NPL ratio	NPL ratio	Delinq Ratio	Delinq Ratio
State absolute discretion	0.05		0.45	
	(0.39)		(0.54)	
State sd direction discretion		0.04		0.41
		(0.32)		(0.44)
Observations	14,229	14,220	14,229	14,220
R-squared	0.24	0.23	0.28	0.28
Quarter FE	Yes	Yes	Yes	Yes
Lagged Bank Controls	Yes	Yes	Yes	Yes

Table 12: Does discretion lead to better predictions of future outcomes?

This table explores the extent to which the discretionary component in ratings is able to predict future bank ratings and near-term loan performance, and how the predictive power varies with examiner-level absolute discretion. We measure the discretionary component of ratings as each exam's directional discretion, i.e., the actual composite rating minus the predicted rating based upon bank observable hard information. In Column 1, the dependent variable is the rating in the next exam, usually 4 quarters away. The dependent variable for Column 2 is whether a bank is considered a troubled bank, an indicator for whether the bank's rating in the next exam is unsatisfactory (≥ 3). In Columns 3 and 4, the dependent variables are changes in the performance of existing loans in the quarter after the exam relative to the quarter before the exam. Panel A presents pooled results across all observations while Panel B shows the relation separately by three terciles for examiner-level average absolute discretion, measured excluding the current observation. T-statistics are in parentheses and represent the partial explanatory power of each rating. The bottom row of Panel B reports p-values testing whether the coefficient for *resid rating x low disc* is equal to the coefficient for *resid rating x high disc*. $*$, $**$, and $***$ denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A:	(1)	(2)	(3)	(4)
	Next Rating	Troubled Bank	NPL Ratio	Delinq Ratio
Exam-level directional discretion	0.373***	0.026***	0.183***	0.266***
	(0.021)	(0.004)	(0.024)	(0.034)
Observations	13,791	13,791	14,555	14,555
R-squared	0.401	0.302	0.266	0.285
Location-quarter FE	Yes	Yes	Yes	Yes

Panel B:	(1)	(2)	(3)	(4)
	Next Rating	Troubled Bank	NPL Ratio	Delinq Ratio
Exam-level dir disc × Low abs disc (LO)	0.334***	0.016**	0.108***	0.140**
	(8.363)	(0.007)	(2.697)	(2.368)
Exam-level dir disc × Med abs disc (LO)	0.374***	0.029***	0.217***	0.327***
	(10.312)	(0.007)	(5.234)	(5.311)
Exam-level dir disc × High abs disc (LO)	0.401***	0.030***	0.205***	0.303***
	(10.452)	(0.009)	(4.817)	(4.926)
Observations	13,791	13,791	14,555	14,555
R-squared	0.401	0.303	0.267	0.285
Location-quarter FE	Yes	Yes	Yes	Yes
Low abs disc = High abs disc (p-value)	0.245	0.244	0.121	0.072
Med abs disc = High abs disc (p-value)	0.498	0.955	0.635	0.676

Table 13: Predicting Troubled Banks - Comparing the power of actual versus counterfactual ratings

This table compares the predictive power of three counterfactual ratings with the actual rating. Panel A reports correlations between the actual and three counterfactual ratings. The predicted rating is the predicted value obtained from regressing actual ratings on bank observable hard information. The reweighted composite rating uses component weights chosen as the set of weights that provide the best estimate of the bank's actual composite rating in the next year (estimated from a regression of the bank's future rating on current component ratings, subject to the constraint that the coefficients sum to one). The counterfactual weights for components C, A, M, E, L, and S are 0.120, 0.199, 0.291, 0.130, 0.145, and 0.115, respectively. The equal weight rating is the composite rating where all components are equally weighted. Panel B compares the ability of actual ratings and counterfactual ratings to predict proxies of future poor (i.e., bottom decile) bank health indicators. Outcomes are measured as an indicator for whether the bank's rating in the following year is unsatisfactory (≥3) (Column 1), whether the change in the non-performing loan ratio in the quarter after the exam relative to the quarter before the exam is roughly in the top 10% of the sample (Column 2), and whether the change in the loan delinquency ratio in the quarter after the exam relative to the quarter before the exam is in the top 10% of the sample (Column 3).

Panel A: Correlations between actual and counterfactual ratings

Correlations	Composite rating	Predicted rating	Reweighted ratings	Equal weight rating
Composite rating	1.00			
Predicted rating	0.87	1.00		
Reweighted rating	0.94	0.85	1.00	
Equal weight rating	0.93	0.85	0.99	1.00

Panel B: Area under the curve (AUC) of actual and counterfactual ratings

	(1)	(2)	(3)
	Troubled Bank	High NPL	High Delinquency
Composite rating	0.8654	0.6468	0.6198
Predicted rating	0.8560	0.6381	0.6145
Reweighted rating	0.8985	0.6724	0.6352
Equal weight rating	0.8942	0.6745	0.6231

For Online Publication

Appendix Table 1: Variation in discretion, beyond federal versus state

Agarwal et al. (2014) show that state examiners assign more lenient ratings compared to federal examiners. This table explores variation in examiner discretion beyond variation associated with state and federal agency differences. Panels A through D report the estimates of directional and absolute discretion from Table 4, separately for the state and federal samples. The sum of the sample sizes for the federal sample and state sample does not equal the sample size for the full sample, because a small number of examiners who have worked for both federal and state agencies are excluded from the subsample analysis.

Panel A: Exam-level, federal sample

Discretion type	Pred rating	Obs	Mean	S.D.	p10	p25	p50	p75	p90
Directional discretion	Non-integer	8,151	0.04	0.32	-0.35	-0.15	0.00	0.20	0.45
Directional discretion	Integer	8,151	0.03	0.36	0.00	0.00	0.00	0.00	0.00
Absolute discretion	Non-integer	8,151	0.24	0.22	0.01	0.06	0.18	0.35	0.54
Absolute discretion	Integer	8,151	0.13	0.33	0.00	0.00	0.00	0.00	1.00

Panel B: Exam-level, state sample

Discretion type	Pred rating	Obs	Mean	S.D.	p10	p25	p50	p75	p90
Directional discretion	Non-integer	6,528	-0.04	0.32	-0.47	-0.23	-0.01	0.13	0.34
Directional discretion	Integer	6,528	-0.04	0.36	0.00	0.00	0.00	0.00	0.00
Absolute discretion	Non-integer	6,528	0.24	0.22	0.01	0.07	0.18	0.35	0.55
Absolute discretion	Integer	6,528	0.13	0.34	0.00	0.00	0.00	0.00	1.00

Panel C: Examiner-level, federal sample

Discretion type	Pred rating	Obs	Mean	S.D.	p10	p25	p50	p75	p90
Directional discretion	Non-integer	1,273	0.03	0.16	-0.15	-0.07	0.02	0.12	0.23
Directional discretion	Integer	1,273	0.03	0.18	-0.14	0.00	0.00	0.06	0.25
Absolute discretion	Non-integer	1,273	0.23	0.12	0.08	0.15	0.22	0.30	0.38
Absolute discretion	Integer	1,273	0.12	0.17	0.00	0.00	0.00	0.20	0.33

Panel D: Examiner-level, state sample

Discretion type	Pred rating	Obs	Mean	S.D.	p10	p25	p50	p75	p90
Directional discretion	Non-integer	1,134	-0.04	0.17	-0.26	-0.14	-0.03	0.06	0.16
Directional discretion	Integer	1,134	-0.04	0.18	-0.27	-0.06	0.00	0.00	0.11
Absolute discretion	Non-integer	1,134	0.23	0.12	0.06	0.14	0.22	0.31	0.39
Absolute discretion	Integer	1,134	0.12	0.17	0.00	0.00	0.00	0.20	0.40

Appendix Table 2: Directional discretion and absolute discretion, with flexible control variables for bank observables

This table replicates Table 4, with exam-level discretion calculated as the residual from a regression of CAMELS ratings regressed on all bank observable variables listed in Panel D of Table 1, each variable squared, as well as all two-way interactions between the variables.

Panel A: Exam-level

Discretion type	Pred rating	Obs	Mean	S.D.	p10	p25	p50	p75	p90
Directional discretion	Non-integer	14,679	0.00	0.32	-0.40	-0.18	0.00	0.17	0.39
Directional discretion	Integer	14,679	0	0.34	0	0	0	0	0
Absolute discretion	Non-integer	14,679	0	0.22	0.001	0.065	0.17	0.34	0.54
Absolute discretion	Integer	14,679	0.12	0.33	0	0	0	0	1

Panel B: Examiner-level

Discretion type	Pred rating	Obs	Mean	S.D.	p10	p25	p50	p75	p90
Directional discretion	Non-integer	2,407	-0	0.16	-0.21	-0.095	0.00	0.09	0.19
Directional discretion	Integer	2,407	0	0.18	-0.2	0	0	0	0.2
Absolute discretion	Non-integer	2,407	0.22	0.12	0.723	0.136	0.21	0.29	0.38
Absolute discretion	Integer	2,407	0.11	0.17	0	0	0	0.2	0.33

Appendix Table 3: Causal impact of discretion in ratings, with flexible control variables for bank observables

This table replicates Table 10 with more flexible control variables for bank observables. Bank controls consist of all bank observable variables listed in Panel D of Table 1, each variable squared, as well as all two-way interactions between the variables. Standard errors in parentheses are adjusted for the two-step instrumental variables procedure and clustered by examiner. $*$, $**$, and $***$ denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Full sample

	(1)	(2)	(3)
Future bank outcome	Next rating	Tier1 ratio - 1yr	Loan growth - 1yr
Pred composite rating	-0.580*	0.865*	-11.341***
	(0.299)	(0.417)	(4.206)
Observations	10,625	13,033	13,040
R-squared	-0.288	0.345	0.049
Bank controls	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes
Location-quarter FE	Yes	Yes	Yes

Panel B: Long run impact

	(1)	(2)	(3)	(4)
Future bank outcome	Tier1 ratio - 2yr	Loan growth - 2yr	Tier1 ratio - 3yr	Loan growth - 3yr
Pred composite rating	1.071**	-13.092*	0.901*	-24.582***
	(0.518)	(7.151)	(0.526)	(9.171)
Observations	12,298	12,304	11,709	11,712
R-squared	0.122	0.083	0.044	0.022
Bank controls	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes
Location-quarter FE	Yes	Yes	Yes	Yes

Panel C: Auxiliary outcomes

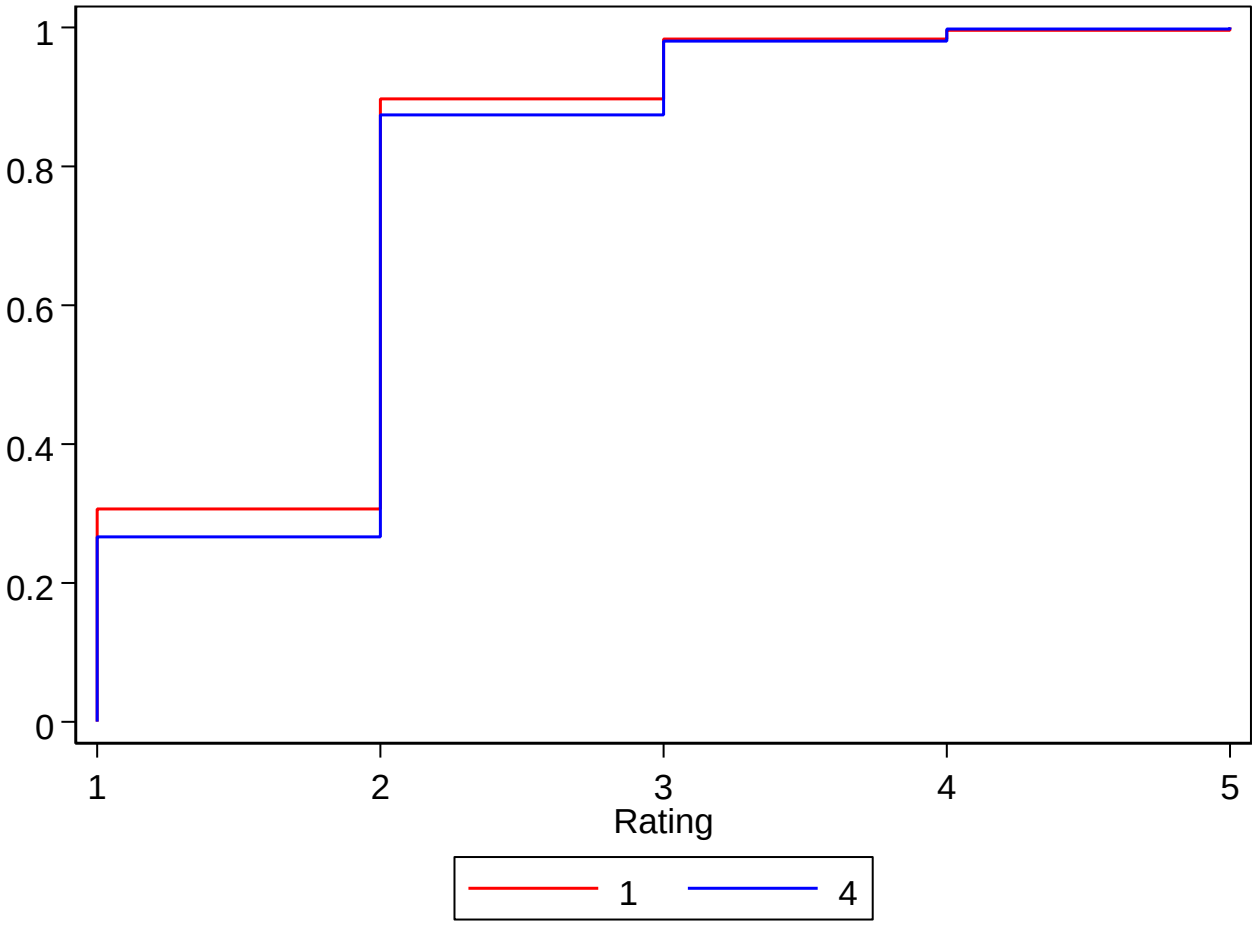
	(1)	(2)
Future bank outcome	NPL ratio	Delinq ratio
Pred composite rating	0.344	0.317
	(0.415)	(0.562)
Observations	13,032	13,032
R-squared	0.213	0.178
Bank controls	Yes	Yes
Bank FE	Yes	Yes
Location-quarter FE	Yes	Yes

Appendix Table 4: Discretion and examiner experience

This table shows how directional discretion and absolute discretion, measured at the exam level, varies with examiner experience, as measured by the log of the number of exams conducted so far by the examiner within our data sample. Standard errors in parentheses are clustered by examiner. $*$, $**$, and $***$ denote statistical significance at the 10%, 5%, and 1% levels, respectively.

	(1)	(2)	(3)	(4)
Exam-level discretion	Directional		Absolute	
Log(Number of exams)	0.007**	0.005	0.016***	0.008***
	(0.004)	(0.005)	(0.002)	(0.003)
Observations	14,679	14,679	14,679	14,679
R-squared	0.000	0.172	0.004	0.199
Examiner FE	No	Yes	No	Yes

Appendix Figure A1: Monotonicity



Following Angrist and Imbens (1994), we test an implication of the monotonicity assumption underlying our instrumental variables analysis. We show that the cumulative distribution function (CDF) of the CAMELS ratings decision for high values of the instrument stochastically dominates the CDF of the CAMELS ratings decision for low values of the instrument. The red and blue lines represent the CDF for the CAMELS rating of the current exam, for examiners with leave-out-mean directional bias in the top (4) and bottom (1) quartiles, respectively.

Appendix: Shrinkage

We account for measurement error in the estimates of examiner-level discretion reported in Table 4. In finite panels, examiner averages are estimated with error, leading to upward bias in the estimate of the variance of discretion across examiners.

We apply an Empirical Bayes shrinkage estimator to estimate the variance of the true magnitude of examiner directional and absolute. Following the methods developed in Morris (1983), with applications provided in Chandra et al. (2016), we assume the estimated examiner directional discretion $\hat{u}_j$ consists of the true examiner directional discretion u_j plus an additive error term ε_j. A similar shrinkage procedure is applied to estimates of examiner-level absolute discretion.

$$\hat{u}_j = u_j + \varepsilon_j.$$

To recover the variance of u_j, we use the "ebayes" Stata package developed by Adam Scarny (`http://sacarny.com/wp-content/uploads/2015/08/ebayes.ado`). For a full description of the assumptions underlying the procedure, we refer the reader to Online Appendix C of Chandra et al. (2016).

Appendix Table 5: Standard Deviation of Examiner Fixed Effects

This table shows the standard deviation of examiner-level directional and absolute discretion in the CAMELS rating and its components. The standard deviation reflects variation in discretion across examiners. For each measure, we present the unshrunken standard deviation and the Empirical Bayes-adjusted standard deviation. The Empirical Bayes-adjusted estimates account for estimation error in finite samples.

	Composite	**C**	**A**	**M**	**E**	**L**	**S**
Directional Discretion							
- Unshrunken	0.183	0.175	0.207	0.207	0.219	0.184	0.185
- Bayes	0.149	0.142	0.168	0.167	0.181	0.151	0.150
Absolute Discretion							
- Unshrunken	0.123	0.121	0.142	0.137	0.153	0.131	0.127
- Bayes	0.092	0.091	0.104	0.101	0.116	0.097	0.094

Appendix: Simulation

In this section, we estimate the fraction of banks that receive a higher or lower rating due to examiner discretion. We focus on banks that would have received a healthy rating of 2 (the modal rating within our sample), absent examiner discretion. We assume that assigned integer ratings are the floor of latent continuous ratings. Absent examiner discretion, banks that would have been assigned a composite rating of 2 are assumed to have latent continuous ratings $2 + x$, where x is distributed according to the following function:

$$f(x) = \left\{ \begin{array}{cc} 0.88254 + .65845 \cdot x & \text{if } x \in [0, 0.5) \\ 1.72960 - 1.03568 \cdot x & \text{if } x \in [0.5, 1] \\ 0 & \text{otherwise} \end{array} \right\}$$

$f(x)$ is a piecewise linear probability density function that is parameterized to match the relative distribution of observations across the CAMELS ratings of 1, 2, and 3 in our data. 28%, 61%, and 9% of bank observations receive CAMELS ratings of 1, 2, and 3 respectively. Note that $f(x)$ peaks in the middle of the unit interval and has lower mass at the left end of the unit interval (banks that are close to receiving a rating of 1) and the lowest mass at the right of the unit interval (banks that are close to receiving a rating of 3). This matches the data showing that the most common rating is 2, and ratings of 3 are less common than ratings of 1. We scale $f(x)$ so that it has an integral of 1.

We model ratings chosen by examiners as *floor(2+x+e)*, where e is drawn from a distribution representing the discretionary component of ratings. We assume that e is drawn from a normal distribution with a mean of zero and a standard deviation of 0.149, equal to the standard deviation of examiner-level directional discretion after applying a shrinkage correction to account for noise. We simulate 1,000,000 random draws for assigned ratings = *floor(2+x+e)*. To estimate the fraction of banks that receive a higher or lower rating due to examiner discretion, we measure the proportion of these random draws with assigned ratings above or below 2. We estimate that 4.2% of banks that would have gotten a rating of 2 absent discretion receive a rating greater than 2 due to discretion. 5.0% of banks that would have gotten a rating of 2 absent discretion receive a rating of 1 due to discretion.

Conditional on being assigned to an examiner in the top quartile of directional discretion (equivalent to drawing an e in the top quartile of its distribution), a bank faces a 13.7% chance of receiving a rating higher than 2.